FROM SINGLE CELLS TO MICROBIAL POPULATIONS: DESENTANGLING MICROBIAL INTERACTIONS AND FUNCTIONS THROUGH INTERGRATED OMICS TECHNIQUES

A Dissertation
Presented to
The Academic Faculty
by

Despina Tsementzi

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Civil and Environmental Engineering

Georgia Institute of Technology
December 2016

FROM SINGLE CELLS TO MICROBIAL POPULATIONS:

DESENTANGLING MICROBIAL INTERACTIONS AND

FUNCTIONS THROUGH INTERGRATED OMICS TECHNIQUES

Approved by

Dr. Konstantinos T. Konstantinidis,
Advisor
School of Civil and Environmental
Engineering
Georgia Institute of Technology

Dr. Frank J. Stewart
School of Biological Sciences
Georgia Institute of Technology

Dr. Frank E. Löffler
School of Civil and Environmental
Engineering
University of Tennessee

Dr. Spyros G. Pavlostathis
School of Civil and Environmental
Engineering
Georgia Institute of Technology

Dr. Jim Spain
School of Civil and Environmental
Engineering
Georgia Institute of Technology

Date approved: November 10 2016

AKNOWLEDGEMENTS

First and foremost I would like than my advisor, Dr. Konstantinidis, for his continuous support and mentoring which made this dissertation possible. I am grateful that I had the opportunity to work with Kostas, not only for him being a great mentor and respectful scientist, but for his patience and trust on me, and all the opportunities that he provided along the way. I thank my committee members, Dr Pavlosthathis, Dr Spain, Dr Löffler and Dr Stewart for the guidance and constructive feedback, as well as for their time and effort towards the completion of this thesis. I would like to thank all the members of Kostas Lab, as well as all my collaborators, for everything that I have learned from them and for making our working environment an inspiring place. I would like to thank the Onassis foundation for their financial support through the first 3 years of my studies. I am forever grateful for my dear friends Luis M. Rodriguez Rojas, Natasha De Leon and Alexandra Meziti for making this journey memorable, and always being there for me. I am grateful to my parents Antonis and Olga Tsementzi, my sister Areti, and brother Giannis for their support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Microbial communities play a central role in global geochemical cycles, environmental engineering systems, and human health. However, how microbial communities perform their *in-situ* activities and what are the biotic (e.g., species interactions) and abiotic controls remain challenging to elucidate and manipulate. The availability of high-throughput sequencing techniques to assess environmental DNA and mRNA levels (functional omics), in combination with new computational tools for analysis of the resulting sequences, provide new opportunities to dissect complex microbial interactions within natural and engineered systems. In this work, we developed a series of new computational tools and pipelines to integrate (microbial) single-cell genomics, metagenomics and metatranscriptomic data from complex ecosystems and recover information on microbial diversity and functional potential. We subsequently applied these techniques to unravel the roles of two highly specialized microbes within complex microbial communities: (a) *Dehaloccoides mcartyi* is the only organism known to date that can perform complete reductive dechlorination of chlorinated contaminants. This critical function for bioremediation applications depends on the poorly understood roles of co-occurring, diverse helper microbial species and complex species interactions. Using integrated omic techniques on natural and engineered mixed communities, we explored the role of community members as well as the strain genomic heterogeneity of *D. mcartyi* in dechlorination activity. (b) The ubiquitous SAR11 bacteria comprise the most abundant organisms in the ocean surface. While SAR11 have been known to only respire oxygen and oxidize small organic molecules, recent findings revealed that these organisms persist in high abundance within extended anoxic water masses formed in oceanic Oxygen Minimum Zones (OMZ). Integrated omic analysis elucidated a previously unrecognized anaerobic metabolism for these cells related to nitrate respiration. These findings link SAR11 to pathways of ocean nitrogen loss, and thus, have important implications for our understanding of OMZs and global nutrient cycling.

CHAPTER 1
INTRODUCTION

Microbial communities play central roles in the global geochemical cycles, environmental engineering systems, as well as in human health and disease. The metabolic activities of microorganisms and their communities have a fundamental impact on the biosphere, from localized micro-niches to the global scale. Understanding the function of microbial communities requires a system-level view to define and quantify the relative importance of the mechanisms that control an individual cellular organism, a population of heterogeneous cells, and an ecological niche [1]. The availability of high-throughput sequencing techniques of environmental DNA and mRNA (functional omics), and the growing reference sequence databases allow the study of otherwise inaccessible (e.g., uncultivated) organisms and genes from a variety of environments. Additionally, the development of new computational tools for sequence analysis enable the study of complex microbial interactions of natural and engineered communities [2,3].

Despite the growing diversity of statistical and computational techniques for functional omics, the rapid growth in data acquisition and reducing sequencing costs have created a lag between the amount of available sequenced datasets and the methods for quantitative assessment and interpretation. Thus, **development of methods for the evaluation and analysis of omics data as well as application of the methods to better understand the activity of the microbiome within natural and engineered systems remain challenging**. In this thesis, a series of new methodological advances, which expand the available toolbox for quantitative analysis of omics data, are presented (Chapters 2-4). These methods were integrated and applied to two different biological systems representing engineered and natural ecosystems to unravel microbial functions and interactions. In Chapter 5, we assess the microbial community interactions within mixed dechlorinating communities, with important implications for bioremediation efforts. Finally, in Chapter 6 we use integrated omic

approaches to unravel new functions of abundant oceanic organisms, with implications for global nutrient fluxes.

1.1 Methodological advances and 'omics techniques

Community-wide metabolic characterization can be performed at different molecular levels, each with different assumptions. Sequencing of total RNA from environmental samples can provide a snapshot of the *in situ* microbial community transcriptional activity and directly assess the gene expression of otherwise inaccessible microbial community members (for instance, see [3]). However, the sequencing processing steps required for the generation of a metatranscriptomic library (RNA extraction, DNAse treatment, rRNA depletion, Amplification, cDNA synthesis, sequencing library construction), together with the short half lives of RNA molecules, might bias the resulting data. Similarly, regulation of bacterial transcription can be highly sensitive to environmental conditions; thus, sample collection protocols that minimize the handling time are required. While the total community DNA libraries (metagenomes) have been shown to be highly reproducible among replicated samples, the reproducibility of metatranscriptomic datasets from environmental samples has remained much more problematic. We recently presented a quantitative evaluation of the reproducibility and variability of different contemporary metatranscriptomic protocols [4]. We showed that the variation observed between technical replicates can be as high as in biological replicates, but proper statistical modeling such as negative bionomial distribution for gene expression data can robustly identify differentially abundant features among replicated samples. To the best of our knowledge, this work represents the first evaluation of technical replication in metatranscriptomics (e.g., see [5]). Additionally, we used the well-replicated metatranscriptomic dataset generated for this study together with companion time-series metagenomics datasets to examine the transcriptional profiles of rare and abundant microbial community members in a freshwater ecosystem. Overall, for most of the bacterial members, the transcriptional activity was proportional to their abundance, e.g., similar RNA/DNA ratios were obtained for different populations. However, significant outliers to this pattern were also detected: Some rare bacteria members showed disproportionally low transcriptional activity, and they typically

belonged to “transient” or "allochthonous" community members, i.e., were not subsequently detected in the time series metagenomes. On the other hand, rare members with disproportionally high activity were detected in high abundances in subsequent time points, indicating that they were persistent members of the community. **Chapter 2** expands on these results by providing a guide for the generation and statistical treatment of metatrascriptomic datasets. New insights with respect to the contribution of rare community members to the community transcriptome and activity that emerged from the application of this methodology to bimonthly samples from a local freshwater planktonic microbial community are also included in this chapter.

Our metatranscriptomics pipeline was also applied to engineered bioreactors that biodegraded an important class of environmental pollutants, the benzakonium chloride disinfectants, in order to identify candidate genes that are involved in the detoxification (biodegradation). The identification and subsequent genetic characterization of those genes (Oh, Kurt, Tsementzi, Weigand, Kim, Hatt, Tandukar, Pavlostathis, Spain, and Konstantinidis, Applied and Environ. Microb. 2014)[6], has important implications in biotechnological applications for cleaning disinfectants in non-target environments, such as the wastewater treatment plants.

An additional degree of complexity in the study of microbial community functions is the existence of intra-population genomic diversity. Studies based on species-level resolution, like the one describe above, implicitly treat species as uniform ecologic units. This simplification is well justified by the typical genetic discontinuity observed in species *in situ* [7,8] as well as the phenotypic distinctiveness considered in the definition of taxonomic ranks [9,10]. However, bacterial strains belonging to the same species can often show extensive gene content variability, while the strain-specific genes can often manifest in drastically different phenotypes (e.g., some pathogenic *vs* non-pathogenic *Escherichia coli* strains share only 70% of their genes) [11]. Typically, it is assumed that the difference in gene content among strains of the same species is due to adaptations to different environmental conditions (e.g., pathogenic *vs* commensal *E. coli* example). It follows that bacterial strains of the same species, which inhabit the same environment, should be more homogenous (clonal) with respect to their genomic content. However, recent studies indicate that sub-species level diversity (i.e., co-existence of multiple different strains in the same sample) might be extensive, and can partially account for ecosystem functioning and resilience [12–14]. For example, in successful treatment cases

of *Clostridium difficile* infections after fecal transplants, the microbial community composition at the gut remains stable at the species level, but the changes within the species (strain level) are highly dynamic, indicating that microbe-host interactions might be mediated by intra-population diversity [15]. Similarly, intra-population diversity of *Accumulibacter phosphatis* has been associated with community resilience and performance stability in an enhance biological phosphorous removal system [16]. Consequently, quantification of intra-population gene content diversity *in situ* (i.e., how much clonal or diverse the genomes of the same species are within the same environment/sample) might help elucidate the underlying gene adaptation and diversity maintenance mechanisms. We recently developed a novel methodology for quantifying intra-population gene-content diversity based on shotgun metagenomes. We hypothesized that the differential coverage of a *de novo* assembly from a single metagenomes is caused by differential representation of the genomic regions in the different cells of the same species. By modeling the observed sequencing depth across the length of a reference genome sequence representing the natural population, we were able to reliable estimate the population core- and pan-genome (i.e., total non-redundant genes within the genomes/cells that make up a population). Application of this methodology to available metagenomes revealed higher intra-population gene content diversity levels for freshwater and marine bacteria, contrasting with genomes from laboratory enrichments, soils and sediments, which usually showed lower diversity levels. In all analyzed natural communities, the median estimated intra-population diversity was lower compared the species-level diversity based on genome culture collections. Extensive gene content diversity was observed for ~20 aquatic populations, mostly *Bacteroides* and *Verrucucomicrobia* members, far exceeding the typical gene content diversity of several named species based on available isolate genome collections, and thus, representing exceptions to the predominant patterns described above. In **Chapter 3** we describe the algorithmic considerations of this method, its implementation and evaluation, and discuss selected cases of natural populations displaying disproportionally high gene-content intra-population diversity. It is also important to note that this work (i.e., Chapter 3) was built upon earlier work on how to best assemble a representative population genome from a complex metagenome (Luo, Tsementzi, Kyrpides, and Konstantinidis, ISME 2012) [17], and what are the differences with this

respect between Illumina and 454 sequencing data (Luo, Tsementzi, Kyrpides, Read, and Konstantinidis, PLoS ONE 2012) [18].

In the work outlined above, we identified functionally and ecologically significant differences between stable and transient members of the rare freshwater community fractions, and discussed the impact of micro-diversity within stable populations in diverse habitats. Distinguishing the sources of transient microorganisms in a community complements their detection and allows the prediction of their potential impacts. For example, transient or dormant animal pathogens may exist in low abundances in the environment, but its main relevance for human activity resides in their potential transmission to animal hosts (including human) and their health impacts [11]. In such cases, it is important to track where the pathogens are coming from (microbial source tracking) in order to limit infection but also for legal purposes. Water quality monitoring is a good example of this issue, in which suffusion of wastewater and other sources of pathogenic microorganisms in water reservoirs or streams may have direct impacts on human health. Traditional methods of water quality monitoring involve the quantification of selective bacteria (e.g., coliforms) or biomarkers, to indirectly quantify potential contamination sources. However, the predictive accuracy of source tracking is not always well established, and improved monitoring tools that can differentiate between pollution sources require a better understanding of bacterial community assembly, functional potential and variability across space and time [19]. In a recent study [20], we aimed to quantify the anthropogenic influence on the microbial community of the Kalamas riverine ecosystem, located in Northeast Greece. The Kalamas river runs though agricultural and urban areas, and thus can serve as a model for detection of anthropogenic influences on dynamic microbial communities. We compared the community assemblages along space throughout the river, and time during 3 different seasons. We developed a computational pipeline to assign potential habitat to the detected bacterial Operational Taxonomic Units (or OTUs; a proxy for species), aiming to assess the contribution of the different inputs in the core microbial community of the river. Our results showed that the river had a much greater input from sewage samples during months of low flow, while typical freshwater bacteria dominated the river stream during the rain reasons and high flow. Additionally, we detected human gut related genes in higher abundance than found in other freshwater aquatic ecosystems, a

possible indication of water contamination from agricultural and/or municipal wastewater. The results, presented in detail in **Chapter 4**, provided a reliable methodology for future studies, including further analyses of watershed ecosystem health, and the identification and development of biomarkers for improved water quality monitoring systems.

1.2 'Omics techniques reveal microbial interactions in engineered consortia

Microbial symbioses are increasingly recognized as naturally frequent and important for community robustness and function, and a key component of efficiency and resilience in engineered microbial consortia [21]. Mixed microbial communities comprising *Dehalococcoides mccartyi* (Dhc) are capable of reductive dechlorination of highly toxic chlorinated compounds, and have central role in bioremediation efforts of chlorinated solvents [22]. *Dhc* are the only dechlorinators known to date that completely detoxify chlorinated ethenes to ethene, encoding unique dehalogenase enzymes that catalyze the reductive dechlorination [23]. However the presence of Dhc in contaminated sites doesn't always result in complete detoxification [24]. The existence of co-occurring, non dechlorinating microbial community members has been implicated to play a significant role in the resilience and bioremediation capability of mixed communities relative to axenic Dhc cultures [25]. However, the specific contributions and relative importance of specific community members remain elusive in most cases. In the work described in **Chapter 5**, we employed a combination of genomics, metagenomics, and metatranscriptomics to characterize metabolic potentials, and disentangle the roles of inter-population interactions in community function within a mixed dechlorinating microbial consortium. Additionally, using the methodology described in Chapter 3 we demonstrate the coexistence of multiple *D. mccartyi* strains in dechlorinating enrichments, which encode different dehalogenase enzymes, potentially driving community resilience and function.

1.3 'Omics techniques reveal surprising metabolic potentials

Bacteria of the SAR11 clade have been isolated from all major ocean regions, and have been characterized as the world's most abundant organism, as they can

comprise -in some cases- up to ½ of all microbial cells in the oxygen-rich surface ocean [26]. Because of their abundance, SAR11 bacteria play a critical role in the marine carbon cycle through the aerobic consumption of dissolved organic matter [27]. All known isolates respire oxygen and oxidize small organic molecules, and SAR11 has not yet been linked to pathways of anaerobic nutrient flux, such as the reductive cycling of nitrogen that occurs in marine oxygen minimum zones (OMZs). OMZs regulate the ocean nitrogen budget as sites where bioavailable nitrogen is lost from the ocean through the anaerobic microbial metabolisms of denitrification and anaerobic ammonium oxidation (anammox) [28]. Together, these processes in OMZs account for up to 50% of total nitrogen lost from the ocean to the atmosphere as N_2 or N_2O gas [29]. Hence, OMZs drive ocean nitrogen content, and therefore productivity, potentially affecting even the climate through the emission of a potent greenhouse gas, N_2O, a bi-product of denitrification and anammox. SAR11 bacteria have been previously identified in high abundance within anoxic water masses of the worlds major OMZs[30]. Given their known metabolism of aerobic oxidation of organic matter, their prevalence in anoxic water is unexpected, while available metagenomics data haven't provided an explanation for this observation. The high diversity of the SAR11 genomes coinciding in the same sample typically prohibits the generation of long *de novo* sequence assemblies, thus linking genes to specific organisms remains challenging for this group. We hypothesized that SAR11 cells harbor adaptations that allow them to thrive in the anoxic OMZ waters, and employed an approach combining single cell genomics, metagenomics, and metatranscriptomics to unravel their metabolic potential [31]. We found that most of the SAR11 cells within the OMZ harbor multiple copies of functional nitrate reductases, which were subsequently validated based on genetic manipulations in the laboratory, and thus have the capacity for anaerobic oxidation of organic matter, coupled to nitrate reduction. We further used quantitative metagenomic and metatranscriptomics datasets to identify the relative contribution of SAR11 cells in the OMZ nitrogen cycle. **Chapter 6** describes the identification and characterization of a new metabolic pathway for the ocean's most abundant organism, directly linking them to the nitrogen cycle within the OMZ. Those findings have broader implications on our understanding of the evolution and adaptability of SAR11 cells, as well as on the way we consider the nutrient fluxes through in the OMZ system: SAR11 cells have the potential to directly utilize organic carbon and

produce nitrate, providing a critical intermediate for the denitrification and anammox pathways, which together account for about half of the nitrogen loss from the ocean.

1.4 REFERENCES

1. Zengler, K. & Palsson, B. O. A road map for the development of community systems (CoSy) biology. *Nat. Rev. Microbiol.* **10,** 366–372 (2012).
2. Segata, N. *et al.* Computational meta'omics for microbial community studies. *Mol. Syst. Biol.* **9,** 666 (2013).
3. Zhou, J. *et al.* High-Throughput Metagenomic Technologies for Complex Microbial Community Analysis: Open and Closed Formats. *mBio* **6,** e02288-14 (2015).
4. Tsementzi, D., Poretsky, R., Rodriguez-R, L. M., Luo, C. & Konstantinidis, K. T. Evaluation of metatranscriptomic protocols and application to the study of freshwater microbial communities. *Environ. Microbiol. Rep.* **6,** 640–655 (2014).
5. Xia, Y. *et al.* Thermophilic microbial cellulose decomposition and methanogenesis pathways recharacterized by metatranscriptomic and metagenomic analysis. *Sci. Rep.* **4,** 6708 (2014).
6. Oh, S. *et al.* Microbial Community Degradation of Widely Used Quaternary Ammonium Disinfectants. *Appl. Environ. Microbiol.* **80,** 5892–5900 (2014).
7. Caro-Quintero, A. & Konstantinidis, K. T. Bacterial species may exist, metagenomics reveal. *Environ. Microbiol.* **14,** 347–355 (2012).
8. Rodriguez-R, L. M. & Konstantinidis, K. T. Estimating coverage in metagenomic data sets and why it matters. *ISME J.* (2014). doi:doi:10.1038/ismej.2014.76
9. *Bergey's Manual® of Systematic Bacteriology*. (Springer New York, 2001).
10. Konstantinidis, K. T. & Stackebrandt, E. in *The Prokaryotes* (eds. Dworkin, M., Falkow, S., Rosenberg, E., Schleifer, K.-H. & Stackebrandt, E.) **1,** 29–57 (Springer-Verlag, 2013).
11. Luo, C. *et al.* Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. *Proc. Natl. Acad. Sci.* **108,** 7200–7205 (2011).
12. Simmons, S. L. *et al.* Population Genomic Analysis of Strain Variation in Leptospirillum Group II Bacteria Involved in Acid Mine Drainage Formation. *PLOS Biol* **6,** e177 (2008).
13. Greenblum, S., Carr, R. & Borenstein, E. Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species. *Cell* **160,** 583–594 (2015).
14. Gruber-Dorninger, C. *et al.* Functionally relevant diversity of closely related Nitrospira in activated sludge. *ISME J.* **9,** 643–655 (2015).
15. Li, S. S. *et al.* Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science* **352,** 586–589 (2016).
16. Wilmes, P. *et al.* Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. *ISME J.* **2,** 853–864 (2008).
17. Luo, C., Tsementzi, D., Kyrpides, N. C. & Konstantinidis, K. T. Individual genome assembly from complex community short-read metagenomic datasets. *ISME J.* **6,** 898–901 (2012).

18. Luo, C., Tsementzi, D., Kyrpides, N., Read, T. & Konstantinidis, K. T. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PloS One* **7,** e30087 (2012).
19. McLellan, S. L. & Eren, A. M. Discovering new indicators of fecal pollution. *Trends Microbiol.* **22,** 697–706 (2014).
20. Meziti, A., Tsementzi, D., Ar. Kormas, K., Karayanni, H. & Konstantinidis, K. T. Anthropogenic effects on bacterial diversity and function along a river-to-estuary gradient in Northwest Greece revealed by metagenomics. *Environ. Microbiol.* n/a-n/a (2016). doi:10.1111/1462-2920.13303
21. Hays, S. G., Patrick, W. G., Ziesack, M., Oxman, N. & Silver, P. A. Better together: engineering and application of microbial symbioses. *Curr. Opin. Biotechnol.* **36,** 40–49 (2015).
22. Taş, N., Van Eekert, M. H. A., De Vos, W. M. & Smidt, H. The little bacteria that can – diversity, genomics and ecophysiology of 'Dehalococcoides' spp. in contaminated environments. *Microb. Biotechnol.* **3,** 389–402 (2010).
23. Löffler, F. E. *et al.* Dehalococcoides mccartyi gen. nov., sp. nov., obligately organohalide-respiring anaerobic bacteria relevant to halogen cycling and bioremediation, belong to a novel bacterial class, Dehalococcoidia classis nov., order Dehalococcoidales ord. nov. and family Dehalococcoidaceae fam. nov., within the phylum Chloroflexi. *Int. J. Syst. Evol. Microbiol.* **63,** 625–635 (2013).
24. Maphosa, F. *et al.* Ecogenomics of microbial communities in bioremediation of chlorinated contaminated sites. *Front. Microbiol.* **3,** (2012).
25. Hug, L. A., Beiko, R. G., Rowe, A. R., Richardson, R. E. & Edwards, E. A. Comparative metagenomics of three Dehalococcoides-containing enrichment cultures: the role of the non-dechlorinating community. *BMC Genomics* **13,** 327 (2012).
26. Salter, I. *et al.* Seasonal dynamics of active SAR11 ecotypes in the oligotrophic Northwest Mediterranean Sea. *ISME J.* **9,** 347–360 (2015).
27. Tripp, H. J. The unique metabolism of SAR11 aquatic bacteria. *J. Microbiol. Seoul Korea* **51,** 147–153 (2013).
28. Lam, P. & Kuypers, M. M. M. Microbial nitrogen cycling processes in oxygen minimum zones. *Annu. Rev. Mar. Sci.* **3,** 317–345 (2011).
29. Codispoti, L. A. *et al.* The oceanic fixed nitrogen and nitrous oxide budgets: Moving targets as we enter the anthropocene? *Sci. Mar.* **65,** 85–105 (2001).
30. Ganesh, S., Parris, D. J., DeLong, E. F. & Stewart, F. J. Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. *ISME J.* **8,** 187–211 (2014).
31. Tsementzi, D. *et al.* SAR11 bacteria linked to ocean anoxia and nitrogen loss. *Nature* **536,** 179–183 (2016).

CHAPTER 2

EVALUATION OF METATRANSCRIPTOMIC PROTOCOLS AND APPLICATION TO THE STUDY OF FRESHWATER MICROBIAL COMMUNITIES

2.1 ABSTRACT

Metatranscriptomics of environmental samples enables the identification of community activities without *a priori* knowledge of taxonomic or functional composition. However, several technical challenges associated with the RNA preparation protocols can affect the relative representation of transcripts and data interpretation. Here, seven replicate metatranscriptomes from planktonic freshwater samples (Lake Lanier, USA) were sequenced to evaluate technical and biological reproducibility of different RNA extraction protocols. Organic versus bead-beating extraction showed significant enrichment for low versus high G + C% mRNA populations respectively. The sequencing data were best modelled by a negative binomial distribution to account for the large technical and biological variation observed. Despite the variation, the transcriptional activities of populations that persisted in year-round metagenomes from the same site consistently showed distinct expression patterns, reflecting different ecologic strategies and allowing us to test prevailing models on the contribution of both rare biosphere and abundant members to community activity. For instance, abundant members of the *Verrucomicrobia* phylum systematically showed low transcriptional activity compared with other abundant taxa. Our results provide a practical guide to the analysis of

metatranscriptomes and advance understanding of the activity and ecology of abundant and rare members of temperate freshwater microbial communities.

2.2 INTRODUCTION

High-throughput sequencing of total community RNA (a.k.a. metatranscriptomics) is becoming a standard molecular tool in microbial ecology and has revealed that the community transcriptome is highly dynamic, varying across environments [1–4], and temporal and spatial scales [5–8]. Metatranscriptomics can potentially assess the activity of every gene in the community and monitor subtle changes in gene expression [3,9]. However, data interpretation largely depends on adequate controls and replication to account for the typically large variation observed among replicates, which is attributable to the inherent complexity of metatranscriptomic protocols [3,10,11]. High reproducibility has been recently reported for the steps following RNA extraction [10–14]; however, little is known about the effect of RNA extraction protocols on the diversity of recovered transcripts. Moreover, the use of biological replicates in environmental studies has been scarce [15], limiting the assessment of reproducibility and sampling effect on the derived conclusions. Although several robust and sophisticated statistical approaches have been described recently for the analysis of RNA sequencing datasets from (mostly) pure isolates [16–21], the lack of experimental data, with adequate replication, restricts the application of these methods on environmental samples.

Despite their inherent limitations, metatranscriptomics have been increasingly applied to explore the *in situ* microbial community gene expression in various environments. Among the aquatic habitats, most studies to date focused on marine and coastal environments [5,7–9,22], and only a few studies were conducted on freshwater ecosystems [4]. Metagenomic and gene amplicon sequencing studies have vastly advanced our knowledge on the diversity, structure, and resilience of the microbial communities of freshwater ecosystems [23–25], but the expression profiles of these communities remain largely underexplored. For instance, endemic freshwater bacterial clades have been identified with molecular techniques, and significant gene content differences have been described for these populations relative to their marine counterparts [24–28,28], but whether or not these genes are expressed and persistent

members of the community show different strategies in terms of gene expression (e.g., high vs. low transcriptional activity) remained unknown.

In this study, we described seven deeply sequenced metatranscriptomic datasets derived from replicated filters generated from the same planktonic sample from Lake Lanier (Atlanta, GA), and sought to provide quantitative insights into the previous issues and the transcriptional activity of individual bacterial populations. Our first goal was to evaluate the effect of two commonly used RNA extraction protocols on the recovered transcripts. Second, we aimed to quantify the variation resulting from biological and technical replicates in order to identify the best statistical models and bioinformatic practices to account for the typically large variation observed and robustly detect differentially expressed genes between protocols. Third, we opted to characterize the functions expressed by high and low abundance members of the community and assess their differences and similarities by querying the derived metatranscriptomic datasets against their population genome sequences recovered from companion, replicated metagenomic datasets from the same sampling time (July 2010). To further validate the population genomes and determine whether a detected population represents a consistently rare or abundant member of the Lake Lanier planktonic microbial community, the abundance pattern of each population was evaluated based on seven metagenomes collected year around (2009-2011) from the same site. These data were also used to test prevailing models on the contribution of low abundance (rare) members to total community activity, *e.g.*, that rare members frequently contribute disproportionally high to community activity [29]. Therefore, the present study not only advances the methods for metatranscriptomics and represents a guide to the bioinformatics analysis of metatranscriptomic data but also provides important new insights into the metabolic activity and complexity of temperate freshwater microbial communities. To the best of our knowledge, the present study is the first metatranscriptomic survey of a typical freshwater lake of the Southeast USA and represents the largest sequencing effort to characterize a freshwater metatranscriptome to date.

2.3 EXPERIMENTAL PROCEDURES

2.3.1 Sample collection and processing

All water samples were collected from Lake Lanier, below the Browns Bridge with a horizontal sampler (Wildco Instruments) at the well-oxygenated depth of 3 m. On July 6, 2010 we collected four independent 4 L samples for RNA and two 20 L samples for DNA extractions. Samples for RNA extraction were filtered (10 minutes time for each sample) and frozen directly on the field while those for DNA were filtered in the laboratory with the same filtration scheme: pre-filtration through a 1.6 µm pore size glass fiber filter (Geotech) and collection on a 0.2 µm polyethersulfone (PES) filter (PALL) for RNA samples or a 0.2 µm sterivex filter (Millipore) for the DNA samples, using a peristaltic pump. Five additional year-round time series samples for DNA isolations were collected and processed with the same protocols, two of which have been previously described [24] (Table A1). The four frozen filters collected for RNA isolation were shattered and split in half by weight. Each half filter was processed separately with either one of two commonly used RNA isolation protocols, or with same protocol (preparation A & B) to assess inter-protocol variations: an enzymatic cell lysis combined with organic RNA extraction method, similar to that used in [30] (OP method) or a mechanical cell lysis followed by RNA purification with the RNeasy kit (Qiagen), as used previously for water samples [31] (BP method). Metatranscriptomic samples (MTR) were named based on the number of the filter they were derived from and the RNA extraction protocol. One half of Filter #5 was further split in half prior to extraction, and one of the resulting aliquots (OP5B) was treated with the newly available at the time RiboZero mRNA enrichment kit. All RNA preparations were DNAse treated, mRNA was enriched by rRNA removal using an enzymatic degradation and a subtractive hybridization treatment, amplified and reverse transcribed as described in detail in Appendix 1, Section A.1.

2.3.2 Functional and taxonomic classifications of cDNA and DNA datasets

cDNA and DNA reads were quality-checked and processed as described in the Appendix 1, Section A.1. DNA reads were assembled as previously described [32]. Genes were predicted from the assembled contigs longer than 500 bp from each of all seven time-series metagenomic datasets using MetaGeneMark [33], and functionally classified

using Blastp (score ≥ 250 bits) against the Cluster of Orthologous Groups (COGs) [34] and the Swiss-Prot [35] protein databases; the latter also provided Gene Ontology terms [36]. The taxonomic origin of contigs was predicted using the MyTaxa pipeline [37].

All abundance estimates and comparisons of functional or taxonomic categories in the cDNA or DNA datasets were based on counts of cDNA or DNA reads mapping on the assembled genes of the metagenomes, respectively, unless noted otherwise. cDNA and DNA reads were mapped against the assembled genes using BLAT [38] with parameters "-tileSize 6 -stepSize 5" (only best match considered, identity over the total read length ≥ 95%, alignment ≥ 80% of the read length). All seven time series metagenomic assemblies were used as a reference for mapping cDNA and DNA reads (July 2010), since some rare taxa at a given time might be better represented in the DNA assembly at another time. Sequencing reads encoding ribosomal proteins were identified with Blastp against the curated ribosomal protein database provided by Yutin and colleagues [39] (at least 60% identity over the whole length of the query sequence, both paired reads mapping to the same taxon). Read counts from paired reads were considered independent events and counted separately in all other cases.

2.3.3 Evaluation of reproducibility and statistical modeling of variation

To assess the level of reproducibility and protocol-specific biases, metatranscriptomic datasets were compared pair-wise at different levels: correlation of transcript abundance of (i) predicted metagenomic genes or (ii) open reading frame (ORF) clusters, the latter directly constructed from the cDNA reads. Additionally, (iii) comparisons at the read level were performed by contrasting the k-mer compositions of the datasets. Second, to determine the degree of technical, biological and inter-protocol variation, the degree of over-dispersion (*i.e.,* excess variation not accounted for when assuming the theoretical Poisson distribution) of the data was assessed by testing the fit of different distributions to the data, i.e., Poisson, over-dispersed Poisson (OD-Poisson), and negative binomial (NB) distributions, as proposed elsewhere [18]. Third, after establishing the best statistical approach to account for the variations (NB), the datasets were compared as different treatments (OP vs BP) at different taxonomic and functional categories, in order to identify differentially abundant features and the effect of extraction

protocols in data interpretaion. Further details are provided in the Appendix A, Section A.1.

2.3.4 Detection of differentially abundant features in DNA or cDNA datasets

All comparisons among replicated metatranscriptomes (OP vs. BP) or metatranscriptomes against their companion metagenomes based on different features (*e.g.*, genes, functional annotation, or taxonomic assignment) were performed using the underlying negative binomial distribution (based on evaluation of variability; see above). The R package DESeq [40] was used with a 0.05 adjusted p-value as a threshold for differential abundance [Benjamini-Hochberg correction; [41]]. Counts were normalized across datasets using the median-based normalization of DESeq, and averaged across replicates (four OP, three RP, two metagenomes) [40]. The average normalized counts obtained were used to compare relative expression ratios between different taxa identified, *i.e.*, $\log_2$ (cDNA/DNA; see below).

2.3.5 Assessment of correlation of abundance with transcriptional activity

The transcriptional activity and abundance of taxa were assessed based on the mapping of reads on assembled metagenomic genes (July 2010 datasets; see above). Taxa were defined at both the genus and species levels. Previously characterized (i.e., known) genera recovered in the metagenomes were identified using MyTaxa's taxonomic assignment of contigs and accounted for a small fraction of the total community (~12% of the total reads mapping on contigs). Contigs from each metagenome were also clustered into species-level groups, which represented both known and unknown taxa, based on their tetra-nucleotide frequencies determined by the 2TIER-binning software [42]. The clustering was performed using Pearson's correlation distances ([1-R]/2) and complete linkage clustering method (R function hclust). A cutoff distance of 0.15 (corresponding to Pearson's R=0.7) was used to group together contigs into species-level groups as previously suggested [43,44]. The linearity and monotonicity of the relationship among cDNA and DNA abundance (average normalized counts; see above) were assessed to identify potential biases in transcriptional activity between rare and abundant taxa. The former was done by calculating the Pearson's correlation between cDNA and DNA abundance, and the latter with the Spearman's correlation between the

DNA abundance and the $\log_2$ of the expression ratios, using built-in functions of the R statistical language (cor).

2.3.6 Recovery of population genomes from time series metagenomes

Individual population genomes were recovered using coverage (abundance) correlation in the seven metagenomic datasets, in addition to the tetra-nucleotide frequency correlation mentioned above, essentially as previously described [45]. Assemblies and gene predictions of the recovered genomes can be found at http://enve-omics.ce.gatech.edu/data/MTR. NCBI SRA numbers for all datasets are provided in Table A1.

Table 2.1. Metatranscriptomic samples for evaluation of reproducibility.
In total four filters were collected (biological replicates) and were split in half; each half was treated with a different extraction protocol or with the same protocol and processed and sequenced independently (prep-A & prep-B; technical replicates). One of the aliquots from filter 5 was further spit in half, resulting in another pair of technical replicates (OP5A, OP5B). OP – enzymatic lysis plus organic extraction protocol; BP - bead-beating plus RNeasy protocol

RNA extraction method	**Organic Extraction (OP)**		**Bead-Beading based (BP)**	
	Prep-A	Prep-B	Prep-A	Prep-B
Filter 2	**OP2^**		**BP2**	
Filter 3			**BP3A**	**BP3B^**
Filter 4	**OP4A**	**OP4B**		
Filter 5	**OP5A**	**OP5B***	**BP5**	

*For the mRNA enrichment step of OP5B, Ribo-Zero Removal (Epicenter) was used instead of MicrobeExpress (Ambion), which was used in all other preparations. ^Datasets were excluded from further analysis low sequencing yield.

2.4 RESULTS AND DISCUSSION

Nine metatranscriptomes were obtained in total, four BP and four OP ones, originating from four different half filters (filters #2, 3, 4, 5; biological replicates), plus an additional OP dataset from filter #5, which was treated with a different rRNA removal protocol (Ribo-Zero rRNA) compared to all other datasets (MICROB*Express*; Fig. A1, Table 2.1). After trimming for low quality reads, two of the datasets (BP3B, OP2) were excluded from further analysis due to large number of low quality reads (>70% of the total). The resulting seven datasets were ~1Gb in size on average, ranging between 0.56 and 10 Gb, pair-end (2X100 bp) reads (Table A1). The OP libraries were systematically larger in size, but there was no apparent difference in RNA yields between the two protocols (Table A2) that could explain that difference in the sequencing yield.

2.4.1 Evaluation of metatranscriptomic protocols

Ribo-Zero rRNA displayed the highest rRNA removal efficiency, with only 9% of the resulting sequences identified as rRNA, as opposed to 26-55% rRNA in all other libraries prepared using MICROB*Express* (Fig. A1), consistent with previous reports on bacterial isolates and human stool samples [11]. After removing rRNA transcripts, a large fraction (17-40%) of the resulting sequences within each library represented rare (low abundance) mRNA transcripts (k-mer mode of one), including the library with the highest sequencing effort (~10Gb; 17% low abundance reads). To determine whether these transcripts represented sequence artifacts or real sequences, their relative abundance across the replicated datasets was evaluated. Rare sequences in an individual dataset were neither rare in all replicates, nor they were always poorly represented in reference databases (Fig. A2, Section A.2). These findings indicated that a large variation in low-abundance transcripts should be accounted for by statistical analysis but not by pre-filtering based on (arbitrary) abundance thresholds, since such transcripts don't necessarily represent sequencing artifacts.

In order to evaluate the reproducibility between replicates and different extraction protocols, the expression profiles of the cDNA datasets were compared pairwise using genes assembled from metagenomes and clusters of ORFs identified directly from

cDNA reads (Appendix A, Section A.2). Based on genes or ORF abundance profiles, datasets from different protocols clustered separately (Fig. 2.1 A-C). Indeed, the variability in technical replicates (same filter, same extraction protocol, done independently) was observed to be almost as large as in biological replicates (different filters, same extraction protocol) (Fig. A3), while the extraction step was responsible for the largest variation (Fig. A3, Appendix A, Section A.2). To further explore whether the variation of the extraction step was due to compositional bias or cell lysis differences between the protocols, we contrasted the 21-mer profiles of the datasets and found no systematic separation between protocols (Fig. 2.1D). A good correlation in gene transcript abundances between protocols was observed for individual genomes (representing individual populations, see genome bins below), but different genomes showed different slopes in regression analysis of the OP vs. BP gene transcript abundance values (Fig. A4). These results indicated that the differences between protocols were most likely due to a lysis bias and not biases in sequence composition. However, it should be mentioned that the datasets obtained showed significant differences in G+C% content between the protocols (Fig. 2.1E) and this likely resulted from the differential lysis of organisms with different genomic G+C%. Hence, the G+C% biases observed here may not be observed in other environments that are populated with organisms characterized by different cell lysis efficiencies and/or G+C% content. To evaluate the effect of the variation observed across replicates in detecting differentially abundant features, the datasets from the two protocols were treated as different treatments and several distributions were applied to the data to determine the one with the best fit. A parametric approach was preferred over non-parametric ones because of the increased statistical power resulting from the underlying distribution of gene expression data. The negative binomial distribution displayed the most realistic fit (Appendix A, Section A.2) and hence, was employed, together with conservative thresholds of significance (p-value adjusted 0.05), to detect differentially abundant features with low false discovery rate.

In order to evaluate how the variability between replicates could affect our interpretations, we compared the different protocols as different treatments using the selected negative binomial distribution. The analysis revealed that reads (Fig. 2.1E), genes (Table A4) and genomes (Table A3, Figure S4) with low and high G+C% content

were significantly more abundant in OP and BP libraries, respectively. The BP method captured more transcripts from members of the *Roseobacter*, *Ruegeria* and *Cyanobacteria* genera. The OP libraries were enriched in *Archaea* (Table A2), and proteins related to viral reproduction (typically low G+C%). Specifically, 23 out of 3388 identified GO categories were identified as differentially abundant in OP compared to BP libraries (p-value adjusted ≤ 0.01, negative binomial test) many of which (n=9) were related to DNA replication or viral reproduction. These findings indicated that different RNA extraction protocols should be considered when targeting high or low G+C% organisms. Despite several significant differences (partly related to G+C% content), most features (>98% of the total) were not differentially abundant between the OP and BP protocols. OP libraries consistently yielded more sequencing reads per unit of input DNA, allowing for increased statistical power [40], and were generated with the same cell lysis method to the one used for the companion metagenomes. Thus, the results obtained with the OP method are preferentially reported below, unless otherwise noted.

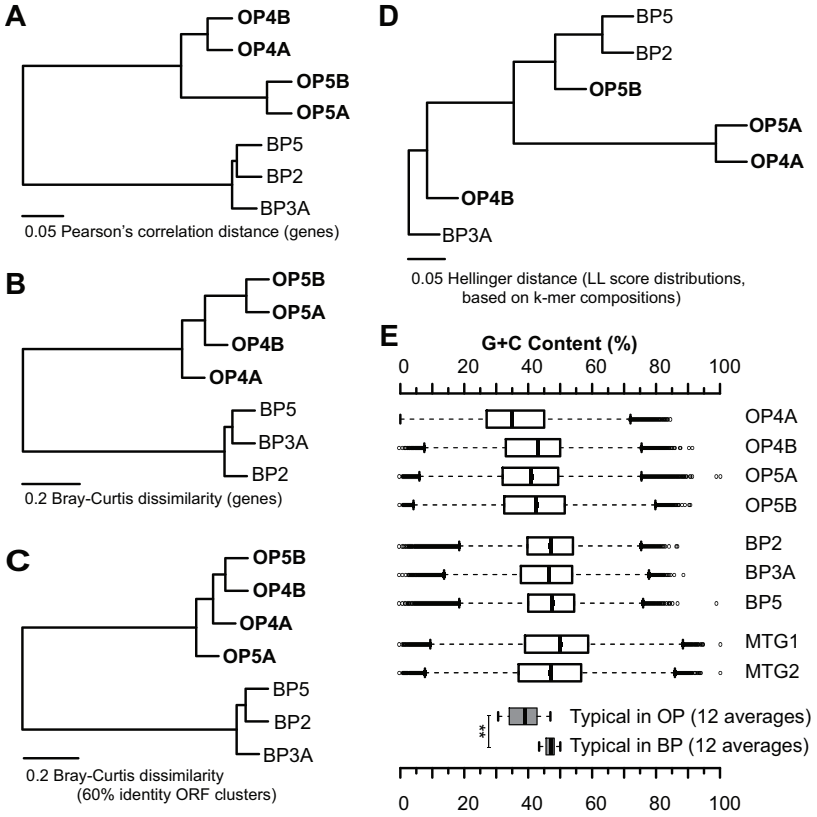


Figure 2.1: Variability and reproducibility among replicated metatranscriptomes. (A-C) The metatranscriptomics datasets clustered by extraction protocol based on the relative abundance of genes (A-B) or ORF 60% clusters (C), using the Pearson's correlation distance (A) or the Bray-curtis dissimilarity index (B-C). (D) A comparison of the distributions of Log-likelihood discriminative scores (LL scores) based on pairwise comparisons of 21-mer composition of the datasets, using Hellinger distances, did not reveal protocol-specific clustering, indicating that compositional biases alone didn't account for the clustering of the datasets by protocol. (E) However, some compositional biases were detected, as shown in the G+C distributions represented as boxplots. The boxes represent the inter-quartile range of the distribution of abundance maxima, traversed by horizontal lines indicating the median. The whiskers extend until the most extreme value within 1.58 times of the inter-quartile range divided by the square root of the number of observations. Values outside this range are represented with circles. The same definition of boxplots is used in other figures as well. In general, the average G+C% content of OP datasets was lower than that of BP datasets (white boxes) and this difference is maintained in the subsets of "typical" reads on each k-mer-based pairwise comparison (grey boxes; one-sided t-test, p-values < 0.01).

2.4.2 Taxon-specific expression levels and physiologies

The taxonomic composition of the microbial community was typical of a freshwater ecosystem [24,25]. The comparison of companion DNA and cDNA datasets revealed a high correspondence of abundance (DNA) and transcriptional activity (cDNA) for the sequences that could be identified at the phylum (R^2=0.97) or the genus (R^2=0.83) levels (Figures 2 and 3A). However, *Verrucomicrobia* showed significant underrepresentation in the cDNA relative to DNA datasets at both taxonomic levels (adjusted p-value 2.1E-11 for phylum, and <0.05 for every detected genus, negative binomial test). Similarly, *Planctomycetes* and the low-abundance phyla *Acidobacteria*, *Deinecococcus-Thermus* and *Synergistetes* contributed significantly less to the community transcripts compared to their DNA abundance (adjusted p-values 1.9E-4, 6.3E-7, 1.2E-6, and 3.8E-3, respectively; Fig. 2.2). *Verrucomicrobia* and *Planctomycetes* also showed low relative expression of ribosomal proteins (Fig. 2.2B; adjusted p-value 6E-4 and 4E-3 respectively), a proposed proxy for relative growth [12,46–49], suggesting that low growth rates and/or distinct cell physiologies likely underlie the low transcriptional activity detected for these two phyla. Alternative explanations such as differential post-translational regulation or mRNA turnover rates remain to be explored. Nevertheless, our results clearly differentiated the *Verrucomicrobia* and *Planctomycetes* populations from those of other taxa, possibly reflecting their distinct physiologies and genomic adaptations.

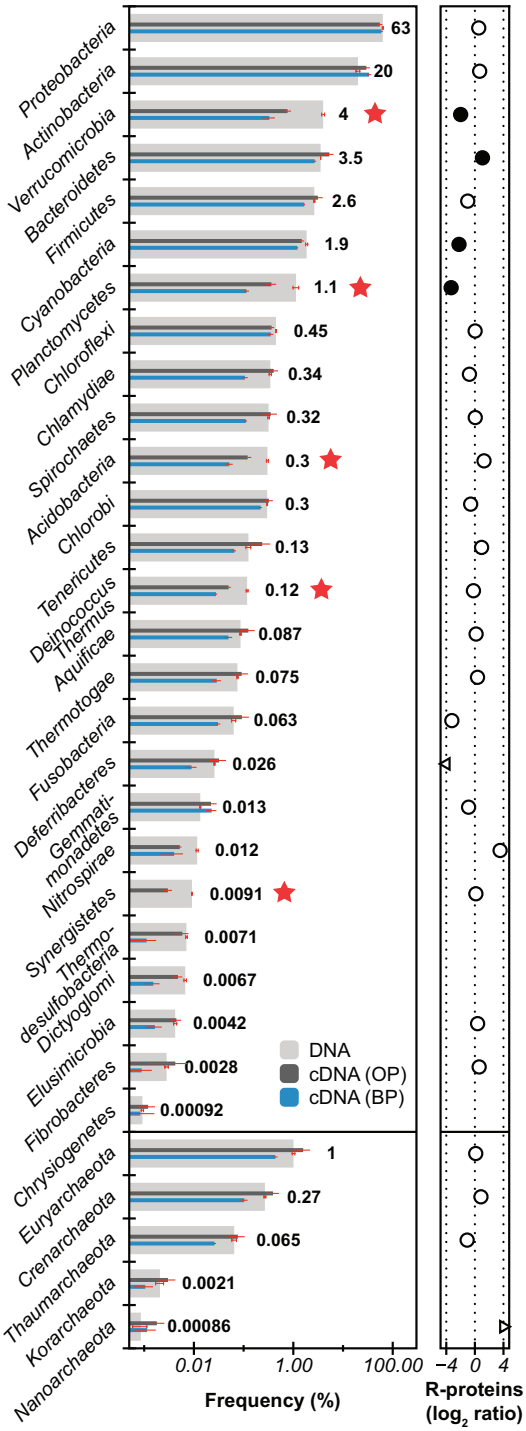


Figure 2.2: Phylum relative abundance and expression levels.

Left panel: Relative abundance of genes (DNA dataset) and their expression levels (OP and BP metatranscriptomes) are shown for each phylum (see figure key; all samples were collected in July 2010). The relative abundance of each gene was estimated as the

average normalized counts (DESeq) per gene divided by the total number of normalized counts mapping to any gene assigned to any phylum. Error bars represent one standard deviation based on replicate datasets. Phyla marked with stars indicate significant underrepresentation in the cDNA datasets (adjusted p-value <0.05). Right panel: Relative expression of ribosomal proteins per phylum were calculated as the $\log_2$ of the ratio of cDNA versus DNA relative abundance values (relative to the total number of reads mapping to ribosomal proteins). Statistically significant low or high ratios are indicated with the black, filled circles (adjusted p-value < 0.05). Empty circles indicate that there was no statistical difference and triangles indicate no detection in DNA or cDNA. All phyla in the figure are represented in the ribosomal database used as reference with the exception of *Thermodesulfobacteria* and *Chrysiogenetes*.

2.4.3 Transcriptional activity of rare community members

To obtain quantitative insights into the contribution of low abundance (rare) members to community activity and begin to test prevailing hypotheses on the functional importance of the rare biosphere [29,50,51], the abundance of a population (DNA datasets) was compared to its transcriptional activity (cDNA datasets). Relative expression ratios (cDNA/DNA) were calculated for all the populations identified at the genus level, which represented approximately 12% of either cDNA or DNA total reads (Fig. A5C). Overall, identified genera contributed to the community transcriptome proportionally to their DNA abundance (Fig. 2.3A, R^2=0.86) and this pattern was consistent across the whole range of population (DNA) abundance (Fig. A6B). This pattern was also reproducible when the analysis was focused on species-like populations based on contig clusters, which recruited >50% of the total reads in each dataset (Fig. A5B). Nevertheless, significant deviations in relative expression ratios were detected for 72 genera (27 with a high expression ratio, 45 low; adjusted p-value < 0.05) that spanned the full range of DNA abundance (Fig. 2.3A and B).

The persistence of these 72 genera in year-around time series metagenomes (2009-2011) was evaluated more closely. Genera were categorized as persistently rare

(n=25), persistently abundant (n=20) or variable (n=27) based on their relative abundance in the time series and using a cutoff for rarity as 0.05% of the DNA reads assigned to the genus level. Abundant members of the community tended to be persistently abundant in the time series data (Fig. 2.3B), independent of their expression levels in the single point metatranscriptome analyzed here. However, rare members with significantly low expression ratios when sampled in July 2010 tended to be persistently rare through time while rare members with significantly high expression ratios were typically found to vary in relative abundance through time, *i.e.*, being abundant in some months and rare in others (Fig. 2.3B and C). These patterns were reproducible when varying the threshold of rarity (0.1, 0.01, 0.05 % of DNA reads) as well as when estimating persistence as the variation in abundance through time without setting a threshold of rarity (data not shown).

It has been previously proposed that rare members of natural communities are disproportionally more active than other members [51–53]. Thus, the data presented here support this model for at least a few members of the community and provided an approach to detect these members based on transcriptional activity and persistence over time. Collectively, 17 out of the total 305 rare genera detected (5.6%) were significantly overrepresented in the cDNA pool, and 9 out of these 17 genera were found to be abundant community members in at least one time point. In contrast, rare taxa with significantly low expression ratios (20 out of the 305, or 6.6%) were typically rare across time (17 out of the 20), indicating that these members of the community likely represent transient or allochthonous populations and rarely contribute to the community transcriptome (Fig. 2.3B and C, Fisher's exact test, p-value 0.032). Indeed when comparing the DNA abundance maxima through time rare members significantly underrepresented in the transcriptome had lower DNA abundance maxima overall (Fig. 2.3C; one-sided t-test, p-value 0.055). Time-series metatranscriptomics will be necessary to further explore the latter hypothesis and study in more detail the ecological strategies of low abundance members. Nevertheless, the replicated datasets presented here indicated that the majority of rare members (>80% of the total) contributed proportionally to their abundance to the transcriptional activity of the community; the few outliers observed represented both constantly rare and low activity members as well as disproportionally highly active and persistent populations. Thus, the two prevailing

models of the rare biosphere are both applicable to freshwater communities and appear to explain different parts of the dynamically changing natural populations.

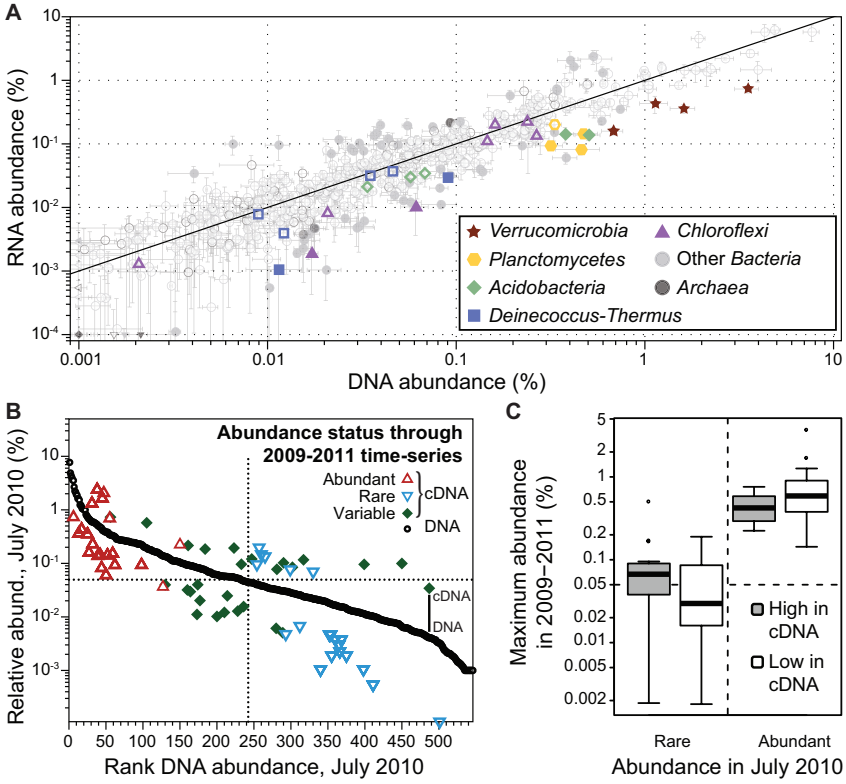


Figure 2.3: Gene expression levels and of rare and abundant populations.

(A) Relative abundance of identified genera in the cDNA (y-axis) versus DNA (x-axis) datasets. Filled symbols represent genera significantly overrepresented in the DNA or cDNA datasets (adjusted p-value <0.05, negative binomial test). Error bars represent one standard deviation based on all OP replicates or the two DNA replicates from July 2010, as estimated by DESeq. Genera discussed in the text are color-coded (figure legend). Note the high correspondence between DNA and cDNA abundances for the majority of the genera in both rare and abundant DNA (cell) fractions. (B) DNA rank abundance curve of identified genera (black) and cDNA relative abundance of selected genera that were significantly over- or under-represented in the transciptome (n=72), color-coded based on their abundance pattern in the time series metagenomes as follows: persistently abundant (red; always > 0.1% in DNA abundance), persistently rare (blue; always < 0.1%) and variable (green). (C) Distribution of DNA abundance maxima in the six time-series metagenomes for identified genera categorized as rare or

abundant based on their abundance in the July 2010 metagenome that were significantly over- (grey box plots) or under-represented (white) in the metatranscriptomic datasets. Note that rare genera identified as having significantly low transcriptional activity (white in C) are persistently rare through time (▽ in B), while rare genera with high transcriptional activity (grey in C) are typically found in the abundant fraction at some time points (◆ in B). The boxplots are defined as in Fig. 2.1. Relative abundance values are based on the fraction of the total reads with taxonomic annotations (phylum level).

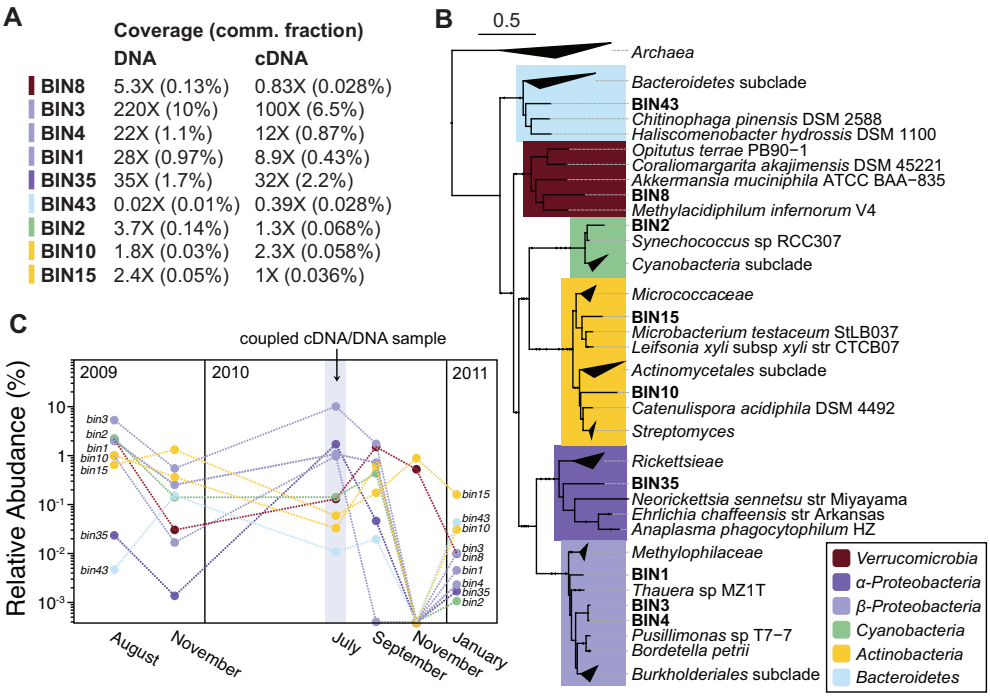


Figure 2.4: Relative abundance and expression level of recovered population genomes.

(A) Genome coverage and average relative abundance (% of total reads) in the DNA (MTG1) and cDNA (OP5A) datasets from July 2010. (B) Maximum likelihood phylogenetic tree of recovered population genomes (BINs) and selected reference bacterial and archaeal genomes based on the concatenated alignment of 30 single copy genes. Tree scale represents substitutions per site. Collapsed clades are represented with triangles where the left vertex represents the common ancestor, and the upper and

lower vertices represent the longest and shortest distance to leaf nodes, respectively. (C) Relative DNA abundance of recovered genomes across the six time series metagenomes, defined as the fraction of the total reads of the dataset mapping to each recovered genome.

2.4.4 Highly expressed functions in the community

As expected, the community transcriptome was dominated by housekeeping functions related to amino acid metabolism, energy production, translation, and replication. Genes related to oxidative phosphorylation and aerobic respiration (oxidoreductases, cytochromes, and ATP synthases), chaperones, topoisomerases, and nitrogen metabolism (glutamine and glutamate synthase, ammonia permeases) as well as hypothetical and poorly characterized genes were among the most highly expressed. Overall, it appeared that the transcriptome was slightly biased towards energy metabolism and protein synthesis as opposed to growth-related categories such as lipid transport, DNA replication, and cell wall synthesis. Similar trends have been observed before in transcriptomes from marine and freshwater ecosystems [4,22,54], and are probably attributable to the fact that microbes typically devote transcriptional activity to the acquisition of nutrients and energy production during daylight and growth at night [4,7,8]; all our samples were collected around noon. Interestingly, genes previously identified to be enriched in metagenomes from freshwater relative to marine ecosystems [24,55], including various potassium transporters, flagellar and viral related proteins, were also found to be among the most enriched transcripts, indicating that these functions typify freshwater communities.

2.4.5 Persistent populations and their expression profiles

Nine phylogenetically diverse population genomes (BINs, Fig. A7), representing persistent, uncultured members of the community and showing a range of *in situ* abundances, were recovered (Fig. 2.4 A and B). Three of the genomes represented members of the *Betaproteobacteria* and were among the most abundant members of the community, recruiting up to 7% of the total DNA reads in July 2010. The recovered

genomes also included representatives of *Alphaproteobacteria*, *Bacteroidetes*, *Actinobacteria*, and *Verrucomicrobia*, and a close relative of marine *Synechococcus* (*Cyanobacteria*). Most transcripts expressed by these groups were devoted to housekeeping genes and genes of hypothetical or unknown function, generally reflecting the patterns observed at the community level (Fig. 2.5). A variety of genes related to oxidative stress response and DNA repair were frequently among the most highly expressed functions, including peroxiredoxins, catalases, and rubrerythrin, reflecting the strong selective pressure of sunlight in the photic surface layer of Lake Lanier. To evaluate ecological specialization among these genomes, the genes expressed were compared among the genomes to identify functions significantly enriched in one or a few genomes relative to the remaining ones or the community average. The most important trends observed are highlighted below.

The low-expression population of Verrucomicrobia (BIN8). Similar to the trend observed above for most members of the *Verrucomicrobia* phylum, BIN8 displayed relatively low transcriptional activity (Fig. 2.4A) and particularly low expression of ribosomal proteins (Fig. 2.5). A predicted proteorhodopsin protein was the eighth most highly expressed transcript in this genome, indicating that photoheterotrophic energy conservation is likely important for this population. Predicted rhodopsin proteins have been previously identified in freshwater *Verrucomicrobia* populations [56], but their *in situ* activity has not been studied. Other highly expressed genes were related to unknown or poorly characterized proteins, as well as several peptidases (mostly serine proteases) and a variety of protein translocation subunits. The latter genes are likely involved in cell wall biosynthesis and/or maintenance and might be related to unique proteinaceous cell wall that characterizes *Verrucomicrobia* [57]. The analysis of this population genome draws attention to the underexplored biology of *Verrucomicrobia*, a phylum with very few cultured representatives [57–59], exhibiting unique physiology with respect to gene transcription and/or regulation and cell organization [57].

Betaproteobacteria (BIN1, BIN3, and BIN4): aerobic anoxygenic phototrophic bacteria (AAPB). The *Betaproteobacteria* bins showed the highest coverage in cDNA and DNA datasets; accordingly, an almost complete genome was recovered for these bins. In all three population genomes, highly expressed bacteriochlorophyll-a

synthetases were detected but no photosystem II genes were identified, indicating that these organisms are AAPB. Other highly expressed genes for those populations were rubrerythrins and proteins related to oxidative stress, phasin proteins, which regulate the synthesis of poly-hydroxyalkanoates (PHAs) inclusions [60], and flagellar proteins. Despite the overall similarities among the transcriptomes of these populations, several differences were also noted such as the distribution of transcripts identified as nutrient transporters (Fig. A8 and S9). Additionally BIN3 and BIN4, which appeared to be close relatives, encoded putative CO oxidation genes and expressed the CO dehydrogenase genes at high levels, indicating that they might employ several strategies to obtain their energy either directly from light or indirectly from CO produced by the photochemical degradation of dissolved organic matter (DOM) [61].

Cyanobacteria (BIN2): Oxygenic photosynthetic population. The clearest signal differentiating this genome from the others was the high expression of oxygenic photosynthesis-related genes as well as phage-related proteins and ABC transporters (Fisher test, adjusted p-value < 1E-15), all of which are consistent with the biology of the closely related genus *Synechococcus* [62,63]. The gene *chrA,* which is involved in chromate resistance by acting as a chemiosmotic pump [64], was also among the top 10% expressed genes. Members of the *Synechoccocus* genus have been shown to exhibit significant resistance to chromate [65], which is typically found in high concentrations in freshwater systems and has been reported to reach levels as high as 20 ppb in Lake Lanier [66,67]. In addition, this genome exhibited the highest detectable transcriptional activity for nitrate transport and expressed a variety of predicted ammonium and phosphorus transporter genes (Fig. A9). Finally, the relatively low expression levels of ribosomal proteins was consistent with the known lifestyle of *Synechoccocus* as abundant and persistent but slow growing organisms [68].

Light-dependent energy conservation in Actinobacteria (BIN10, BIN15) and Alphaproteobacteria (BIN35). The actinobacterial genomes had the lowest coverage in the July 2010 sample and a lower level of genome completeness compared to the other bins. Nevertheless, their gene expression patterns indicated distinct metabolic strategies. High expression of proteorhodopsin for BIN10 and carbon monoxide dehydrogenase subunits for BIN15 were the most distinct functions, both previously reported for

actinobacterial populations [61,69]. On the other hand BIN35 was among the most abundant population genomes but showed the lowest expression ratios of DOC (Dissolved Organic Carbon) transporters, mostly concentrated on lipid transporters (Fig. A8). Among the most highly expressed transcripts from this genome were a predicted proteorhodopsin, transposases, and several phage proteins, possibly indicating an active phage infection of the population at the time of sampling.

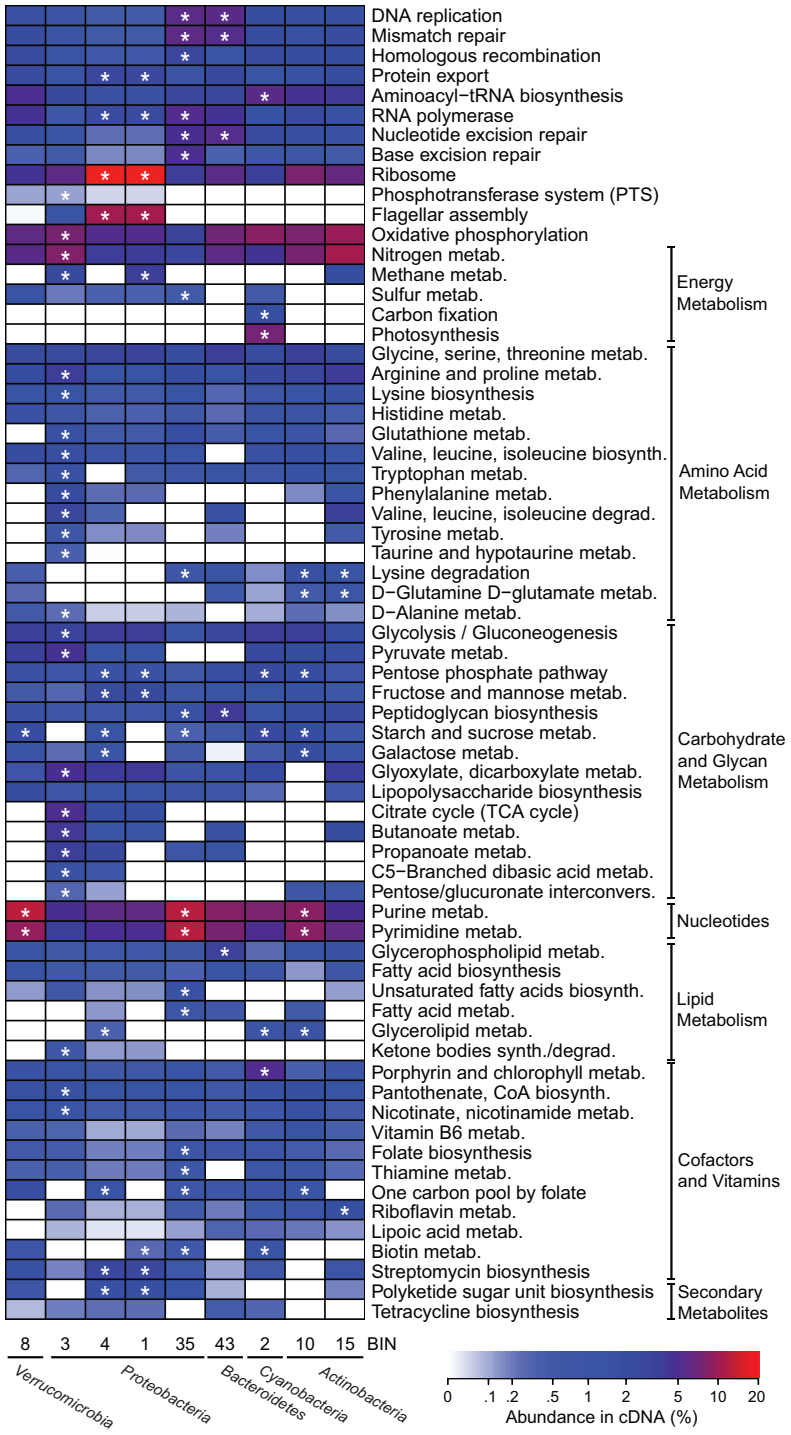


Figure 2.5: Metabolic pathways differentially expressed in recovered genomes. Heatmap of relative expression values, defined as the fraction from the total reads mapping to each genome bin (see scale bar), of KEGG metabolic pathways (rows) for each genome (columns). Asterisks denote metabolic pathways that were significantly

enriched in the corresponding genome compared to the remaining genomes. White cells indicate pathways that were not identified in the corresponding genome based on the MinPath pipeline.

2.5 CONCLUSIONS AND PERSPECTIVES

The present study reported well-replicated metatranscriptomes datasets that allowed assessment of biological and technical variability as well as the advantages of different RNA extraction and isolation protocols. Significant variation has been previously reported for metatranscriptomic libraries from complex environmental microbial communities [3,70], and technical variation resulting from the rRNA removal, mRNA amplification or sample collection steps is considered limited [11,13,14,18]. Our evaluation of libraries from planktonic freshwater ecosystems showed that technical variability could be as high as the variability observed between biological replicates, likely due to several "noise-prone" steps in the underlying protocols. Hence, employing large sequencing efforts with replication and appropriate statistical approaches is critical in increasing the statistical power to detect differentially expressed features. For example, no differentially expressed genes were identified between technical or biological replicates of this study when assuming an underlying negative binomial distribution with local parameters fit, but several significant variations were detected between different extraction protocols. Such variations were partly related to G+C% content, probably caused by differential lysis of the two protocols, and could lead to inaccurate biological conclusions when the samples compared are prepared based on different extraction protocols.

Increased statistical resolution and companion metagenomic datasets allowed us to explore the transcriptomic activity of rare members of the microbial community in Lake Lanier, revealing that most organisms contribute to the community transcriptome proportionally to their cell abundance with a few notable exceptions. Organisms that were consistently rare in the DNA time-series typically contributed comparatively less to the transcriptome, consistent with the hypothesis that such organisms do not represent important players in the community, either because they are dormant cells of allochthonous populations [50]. In contrast, organisms that were present at low abundance at the sampling time but represented persistent members of the community with

relatively high abundance at other sampling times frequently contributed disproportionally high to the community transcriptome. Those taxa that can be found both in rare and abundant fractions of the community through time fit a model in which members of the rare biosphere are under top-down controls and their disproportionally high transcriptional activity, while in small local populations, could lead to their increase in abundance at later time points [29].

Further, the transcriptional profiles of abundant individual populations showed that genes characterizing freshwater microbial communities such as potassium transporters, flagellar related proteins and viral replication proteins, were actively expressed and provided evidence for the ecological and/or metabolic specialization of several of the populations. Most notably, distinct transcriptional patterns were observed for all identified members of the *Verrucomicrobia* phylum, including low transcriptional activity and evidence of slow growth. This phylum exhibits unique structural characteristics such as unique cell compartmentalization and cell wall, but its biology is largely underexplored mainly due to the lack of cultured representatives [57–59,71]. The data presented here provided new insights into the physiology and metabolic role of *Verrucomicrobia* organisms. Similarities among the recovered populations were also observed. For instance, light-dependent energy metabolism was widespread among the most persistent members of the community, either in the form of oxidation of the photochemically produced CO, or with the employment of proteorhodopsins. All together, the results reported here advance our understanding of transcriptional activities and distinct ecological strategies of uncultured members of the Lake Lanier microbial community. These findings likely apply to other temperate freshwater ecosystems as well.

2.6 REFERENCES

1. Frias-Lopez, J. *et al.* Microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 3805–3810 (2008).
2. Urich, T. *et al.* Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PloS One* **3,** e2527 (2008).
3. Gifford, S. M., Sharma, S., Rinta-Kanto, J. M. & Moran, M. A. Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *ISME J.* **5,** 461–472 (2011).
4. Vila-Costa, M., Sharma, S., Moran, M. A. & Casamayor, E. O. Diel gene expression profiles of a phosphorus limited mountain lake using metatranscriptomics. *Environ. Microbiol.* **15,** 1190–1203 (2013).
5. Shi, Y., Tyson, G. W., Eppley, J. M. & DeLong, E. F. Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean. *ISME J.* **5,** 999–1013 (2011).
6. Hewson, I., Poretsky, R. S., Tripp, H. J., Montoya, J. P. & Zehr, J. P. Spatial patterns and light-driven variation of microbial population gene expression in surface waters of the oligotrophic open ocean. *Environ. Microbiol.* **12,** 1940–1956 (2010).
7. Ottesen, E. A. *et al.* Pattern and synchrony of gene expression among sympatric marine microbial populations. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E488-497 (2013).
8. Poretsky, R. S. *et al.* Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ. Microbiol.* **11,** 1358–1375 (2009).
9. Moran, M. A. *et al.* Sizing up metatranscriptomics. *ISME J.* **7,** 237–243 (2013).
10. Stewart, F. J., Ottesen, E. A. & DeLong, E. F. Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J.* **4,** 896–907 (2010).
11. Giannoukos, G. *et al.* Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol.* **13,** R23 (2012).
12. Ottesen, E. A. *et al.* Metatranscriptomic analysis of autonomously collected and preserved marine bacterioplankton. *ISME J.* **5,** 1881–1895 (2011).
13. He, S. *et al.* Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat. Methods* **7,** 807–812 (2010).
14. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* **38,** e131 (2010).
15. Prosser, J. I. Replicate or lie. *Environ. Microbiol.* **12,** 1806–1810 (2010).
16. Fang, Z., Martin, J. & Wang, Z. Statistical methods for identifying differentially expressed genes in RNA-Seq experiments. *Cell Biosci.* **2,** 26 (2012).
17. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11,** 94 (2010).
18. Oberg, A. L., Bot, B. M., Grill, D. E., Poland, G. A. & Therneau, T. M. Technical and biological variance structure in mRNA-Seq data: life in the real world. *BMC Genomics* **13,** 304 (2012).
19. McClure, R. *et al.* Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res.* (2013). doi:10.1093/nar/gkt444

20. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics* **12,** 480 (2011).
21. Parks, D. H. & Beiko, R. G. Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* **26,** 715–721 (2010).
22. Gilbert, J. A. *et al.* Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PloS One* **3,** e3042 (2008).
23. Shade, A. *et al.* Lake microbial communities are resilient after a whole-ecosystem disturbance. *ISME J.* **6,** 2153–2167 (2012).
24. Oh, S. *et al.* Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Appl. Environ. Microbiol.* **77,** 6000–6011 (2011).
25. Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D. & Bertilsson, S. A guide to the natural history of freshwater lake bacteria. *Microbiol. Mol. Biol. Rev. MMBR* **75,** 14–49 (2011).
26. Debroas, D. *et al.* Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (Lac du Bourget--France). *Environ. Microbiol.* **11,** 2412–2424 (2009).
27. Yau, S. *et al.* Metagenomic insights into strategies of carbon conservation and unusual sulfur biogeochemistry in a hypersaline Antarctic lake. *ISME J.* (2013). doi:10.1038/ismej.2013.69
28. Villaescusa, J. A. *et al.* A close link between bacterial community composition and environmental heterogeneity in maritime Antarctic lakes. *Int. Microbiol. Off. J. Span. Soc. Microbiol.* **13,** 67–77 (2010).
29. Pedrós-Alió, C. The rare bacterial biosphere. *Annu. Rev. Mar. Sci.* **4,** 449–466 (2012).
30. Suzuki, M. T. *et al.* Phylogenetic screening of ribosomal RNA gene-containing clones in Bacterial Artificial Chromosome (BAC) libraries from different depths in Monterey Bay. *Microb. Ecol.* **48,** 473–488 (2004).
31. Poretsky, R. S., Sun, S., Mou, X. & Moran, M. A. Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environ. Microbiol.* **12,** 616–627 (2010).
32. Luo, C., Tsementzi, D., Kyrpides, N., Read, T. & Konstantinidis, K. T. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PloS One* **7,** e30087 (2012).
33. Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* **38,** e132 (2010).
34. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4,** 41 (2003).
35. Wu, C. H. *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **34,** D187-191 (2006).
36. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29 (2000).
37. Luo, C., Rodriguez-R, L. M. & Konstantinidis, K. T. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Res.* **42,** e73 (2014).
38. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
39. Yutin, N., Puigbò, P., Koonin, E. V. & Wolf, Y. I. Phylogenomics of prokaryotic ribosomal proteins. *PloS One* **7,** e36972 (2012).

40. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
41. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57,** 289–300 (1995).
42. Saeed, I., Tang, S.-L. & Halgamuge, S. K. Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. *Nucleic Acids Res.* **40,** e34 (2012).
43. Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. & Glöckner, F. O. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.* **6,** 938–947 (2004).
44. Dick, G. J. *et al.* Community-wide analysis of microbial genome sequence signatures. *Genome Biol.* **10,** R85 (2009).
45. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* (2013). doi:10.1038/nbt.2579
46. Wei, Y. *et al.* High-density microarray-mediated gene expression profiling of Escherichia coli. *J. Bacteriol.* **183,** 545–556 (2001).
47. Fazio, A. *et al.* Transcription factor control of growth rate dependent genes in Saccharomyces cerevisiae: a three factor design. *BMC Genomics* **9,** 341 (2008).
48. Hendrickson, E. L. *et al.* Global responses of Methanococcus maripaludis to specific nutrient limitations and growth rate. *J. Bacteriol.* **190,** 2198–2205 (2008).
49. Gifford, S. M., Sharma, S., Booth, M. & Moran, M. A. Expression patterns reveal niche diversification in a marine microbial assemblage. *ISME J.* **7,** 281–298 (2013).
50. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320,** 1034–1039 (2008).
51. Campbell, B. J., Yu, L., Heidelberg, J. F. & Kirchman, D. L. Activity of abundant and rare bacteria in a coastal ocean. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 12776–12781 (2011).
52. Hunt, D. E. *et al.* Relationship between abundance and specific activity of bacterioplankton in open ocean surface waters. *Appl. Environ. Microbiol.* **79,** 177–184 (2013).
53. Jones, S. E. & Lennon, J. T. Dormancy contributes to the maintenance of microbial diversity. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 5881–5886 (2010).
54. Stewart, F. J., Sharma, A. K., Bryant, J. A., Eppley, J. M. & DeLong, E. F. Community transcriptomics reveals universal patterns of protein sequence conservation in natural microbial communities. *Genome Biol.* **12,** R26 (2011).
55. Eiler, A. *et al.* Productivity and salinity structuring of the microplankton revealed by comparative freshwater metagenomics. *Environ. Microbiol.* n/a-n/a (2013). doi:10.1111/1462-2920.12301
56. Martinez-Garcia, M. *et al.* High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. *ISME J.* **6,** 113–123 (2012).
57. Lee, K.-C. *et al.* Phylum Verrucomicrobia representatives share a compartmentalized cell plan with members of bacterial phylum Planctomycetes. *BMC Microbiol.* **9,** 5 (2009).
58. Scheuermayer, M., Gulder, T. A. M., Bringmann, G. & Hentschel, U. Rubritalea marina gen. nov., sp. nov., a marine representative of the phylum 'Verrucomicrobia', isolated from a sponge (Porifera). *Int. J. Syst. Evol. Microbiol.* **56,** 2119–2124 (2006).

59. Matsuzawa, H., Tanaka, Y., Tamaki, H., Kamagata, Y. & Mori, K. Culture-dependent and independent analyses of the microbial communities inhabiting the giant duckweed (Spirodela polyrrhiza) rhizoplane and isolation of a variety of rarely cultivated organisms within the phylum Verrucomicrobia. *Microbes Environ. JSME* **25,** 302–308 (2010).
60. Matsumoto, K. 'ichiro, Matsusaki, H., Taguchi, K., Seki, M. & Doi, Y. Isolation and characterization of polyhydroxyalkanoates inclusions and their associated proteins in Pseudomonas sp. 61-3. *Biomacromolecules* **3,** 787–792 (2002).
61. King, G. M. & Weber, C. F. Distribution, diversity and ecology of aerobic CO-oxidizing bacteria. *Nat. Rev. Microbiol.* **5,** 107–118 (2007).
62. Scanlan, D. J. *et al.* Ecological genomics of marine picocyanobacteria. *Microbiol. Mol. Biol. Rev. MMBR* **73,** 249–299 (2009).
63. Holtman, C. K. *et al.* High-throughput functional analysis of the Synechococcus elongatus PCC 7942 genome. *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes* **12,** 103–115 (2005).
64. Alvarez, A. H., Moreno-Sánchez, R. & Cervantes, C. Chromate efflux by means of the ChrA chromate resistance protein from Pseudomonas aeruginosa. *J. Bacteriol.* **181,** 7398–7400 (1999).
65. Gupta, A., Bhagwat, S. G. & Sainis, J. K. Synechococcus elongatus PCC 7942 is more tolerant to chromate as compared to Synechocystis sp. PCC 6803. *Biometals Int. J. Role Met. Ions Biol. Biochem. Med.* **26,** 309–319 (2013).
66. Brouckaert, B., Amirharajah, A., Zhu, G. & York, T. Heavy metal loading to Lake Lanier from point sources of pollution and urban runoff. in 474–477 (The University of Georgia, 1997).
67. Leigh, D. S. & Gamble, D. W. Survey of nonpoint trace metal inputs to Lake Lanier. in 170–173 (The University of Georgia, 1997).
68. Vaulot, D., Marie, D., Olson, R. J. & Chisholm, S. W. Growth of Prochlorococcus, a Photosynthetic Prokaryote, in the Equatorial Pacific Ocean. *Science* **268,** 1480–1482 (1995).
69. Sharma, A. K., Zhaxybayeva, O., Papke, R. T. & Doolittle, W. F. Actinorhodopsins: proteorhodopsin-like gene sequences found predominantly in non-marine environments. *Environ. Microbiol.* **10,** 1039–1056 (2008).
70. Hollibaugh, J. T., Gifford, S., Sharma, S., Bano, N. & Moran, M. A. Metatranscriptomic analysis of ammonia-oxidizing organisms in an estuarine bacterioplankton assemblage. *ISME J.* **5,** 866–878 (2011).
71. Hedlund, B. P., Gosink, J. J. & Staley, J. T. Verrucomicrobia div. nov., a new division of the bacteria containing three new species of Prosthecobacter. *Antonie Van Leeuwenhoek* **72,** 29–38 (1997).

CHAPTER 3

A METAGENOMIC APPROACH FOR SOURCE TRACKING AND ASSESSMENT OF COMMUNITY ASSEMBLY IN A RIVERINE ECOSYSTEM

Partially reproduced with permission from Alexandra Meziti*, Despina Tsementzi*, Konstantinos Ar. Kormas, Hera Karayanni, Konstantinos T. Konstantinidis. *Environ Microbiol.* 2016 Mar 22. *Equal contribution authors.

3.1 ABSTRACT

Studies assessing the effects of spatial and temporal factors on the taxonomic or functional diversity of bacterioplankton communities in lotic ecosystems are limited, while those combining the effects of anthropogenic perturbations are even more rare. To assess the relative contribution of these dimensions on river water bacterial communities, we applied 16S rRNA gene amplicon and whole-genome shotgun sequencing on samples from the Kalamas River (Northwest Greece), which runs through NATURA-protected sites and recreational areas but also receives urban and industrial, treated and untreated, sewage from the city of Ioannina (150,000 inhabitants) through the Lapsista ditch. Three different locations were sampled, between the exit of the ditch and the estuary, in three different seasons. We found that temporal differences of taxonomic as well as functional diversity were more pronounced than spatial ones. Comparisons of gene diversity with other freshwater and estuarine habitats showed that only February samples resembled freshwater environments. In contrast, taxonomic and functional signals related with human gut bacteria and sewage inputs were far more increased in November, even tens of kilometers downstream of Ioannina city, while the highest signals of soil related and highly heterotrophic communities were detected in May, presumably as the effect of decreased water flow. These findings showed the significance of allochthonous inputs that may have an impact on the formation of

bacterioplankton communities and highlighted the potential of metagenomics for source tracking purposes.

3.2 INTRODUCTION

Streams and rivers constitute a small part of freshwater ecosystems (0.006%) but are critical for a variety of human activities, including sources of drinking water and hydroelectric power as well as sites for recreational and agricultural activities [1]. Additionally, lotic ecosystems are highly dynamic, with constantly changing characteristics influenced from hydrological and climatic factors as well as from anthropogenic and terrestrial inputs. Despite having being considered traditionally as 'passive transporters' of organic matter and nutrients from the land to coastal areas, numerous studies over the last couple decades have revealed their significance in biogeochemical nutrient cycling, temporary storage, and transformation of complex organic compounds of terrestrial origin [2,3]. Although prokaryotic organisms are considered key players in nutrients transformation and energy flow in lotic ecosystems [4–6], our knowledge on bacterioplankton diversity and its response to environmental conditions remains still limited. So far such studies have primarily focused on bacterial community diversity along rivers and streams and its relation with environmental, hydrological and geographical parameters [5,7–12], while only a few studies have examined bacterial communities over time as well [9]. Studies employing shotgun metagenomic profiles of rivers are even more limited and have focused on spatial differences along the river and the effect of land cover [13] or on comparisons with other freshwater ecosystems [8]. Bacterioplankton taxonomic diversity has been studied over space and time only once and based on 16S rRNA gene amplicon sequencing, in streams of Colorado Rocky Mountains, exhibiting pronounced temporal variability with no apparent seasonality [9]. These patterns were driven by intermittent changes of water biogeochemical properties, pointing out the fluctuating character of bacterioplankton communities in streams. However, it remains unclear whether or not these patterns characterize other rivers with different physicochemical and flow dynamics properties.

Another important, but under-studied aspect of surveying bacterioplankton diversity in lotic ecosystems relates to the potential to identify the impact of human

activities. The discharge of treated and untreated sewage from human and industrial activities in rivers and streams can influence bacterioplankton diversity with the augmentation of pathogenic and gut related bacteria as well as antibiotic resistance genes (ARGs) and transposable elements [10,14,15], which also influence water quality and usage. 16S rRNA gene based surveys do not typically provide the resolution required to discriminate between human/animal-adapted microbial populations relative to their environmental close relatives [16]. For this reason, whole genome approaches could contribute towards a more realistic concept on the issue of the anthropogenic footprint in lotic systems.

In this study, we concomitantly analyzed for the first time the taxonomic and functional bacterioplankton diversity in a temperate river along space and time. Kalamas river (Northwest Greece) has a length of about 115 km and a mean annual flow of 54 m^3 s^{-1} [17]. Before discharging to the Ionian Sea, Kalamas River runs through agricultural and natural protected areas ('NATURA' sites, established in 2000). Approximately 85 km upstream of its estuary, Kalamas River receives the overflow of the eutrophic lake Pamvotis of Ioannina city, as well as treated and untreated sewage from this city (approx. 150,000 inhabitants) and from industrial and agricultural units of the area, through an artificial ditch (Lapsista) (Fig. 3.1A). Previous studies on Kalamas River have mainly focused on the concentration of pesticides [18] and the assessment of water and habitat quality [19,20], revealing seasonal differentiations mostly related with water flow (low in the summertime, dry season; high during fall and wintertime rainy seasons). Previous studies on microorganisms have investigated the presence of pathogens such as *Listeria* and *Salmonella* and their susceptibility to antibiotics in temporal and seasonal scale [21], but whole-community microbial diversity has never been studied to date. Here we used 16S rRNA gene amplicon 454-pyrosequencing and whole-genome shotgun (WGS) Illumina sequencing to investigate taxonomic and functional bacterioplankton diversity respectively at the whole community level along the river. Three different sites were selected downstream of the Lapsista ditch exit to Kalamas River and were sampled at three different seasons, characterized by different water flows, in order to observe spatial and temporal changes. We focused on tracking signals that could predict the impact of allochthonous inputs in Kalamas River water bacterial communities and thus, advance our understanding of the highly variable character of the river and the effects of fluctuating environmental factors.

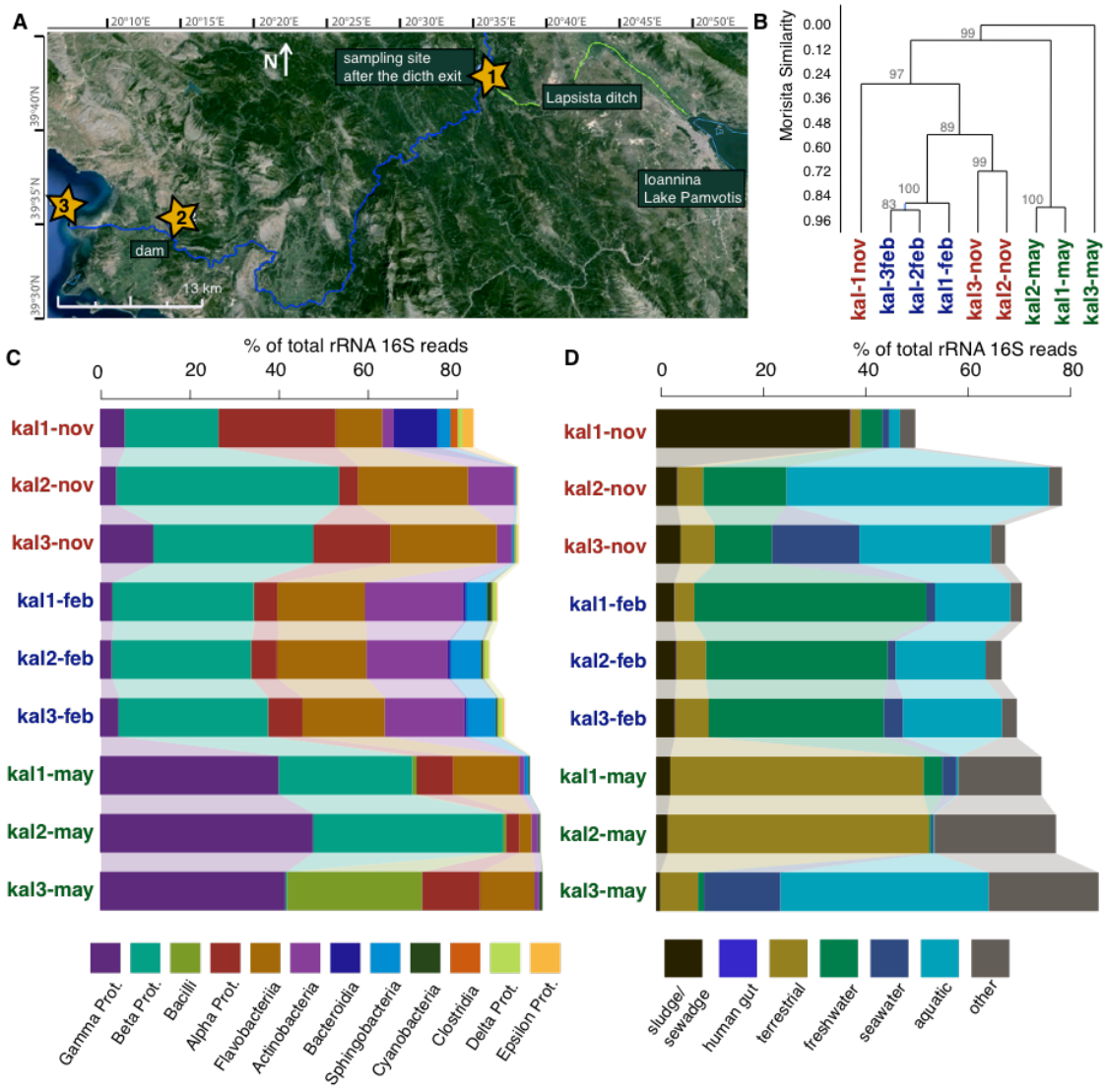


Figure 3.1: Composition of Kalamas River microbial communities.

(A) Location of sampling sites across the river (blue), ditch (green). The land use map was constructed with the ArcGIS platform, World Land Cover30m BaseVue2013 layer (ESRI 2016. ArcGIS platform. Redlands). Deciduous/Evergreen forest: trees with >3m height. Shrub/Scrub: Woody vegetation <3m in height. Grassland: Herbaceous grasses. Agriculture: cultivated crop lands. Water: All water bodies greater than 0.08 hectares. High and medium Density Urban: Areas with over 70%, or between 30-70% of constructed materials that are a minimum of 60 meters wide (asphalt, concrete, buildings, etc.). (B) Clustering of Kalamas samples based on OTU abundance profiles. (C) Taxonomic distributions 16S rRNA amplicon sequence datasets. (D) Habitat assignment

of 16S rRNA sequences based on similarities (>99%) to reference sequences (SILVA119).

3.3 MATERIALS AND METHODS

3.3.1 Sample and metadata collection

Samples were collected in November (Nov) 2012, February (Feb), and May (May) 2013 from three sites along Kalamas River (Fig. 3.1A). The first site (kal1) was selected for its presumably high influence by the contents of the Lapsista ditch. The second site (kal2), just before the hydroelectric dam and 68.6 Km from the Lapsista ditch exit, was selected in order to assess the recovery of bacterial communities downstream of the ditch exit and the influence of water flow rate. Finally, the third site (kal3), located at the estuary, was selected to evaluate how the bacterioplankton diversity is affected by freshwater/saltwater transition. Surface water samples (5L) were collected from the stream bank in sterile plastic carboys and were processed within 2h. Samples were initially filtered through 180μm nylon mesh for the removal of large particles and metazoa, followed by low vacuum filtration (<150 mmHg) on 0.2 μm isopore filters (Sartorius Stedim Biotech, Germany). Temperature, salinity and pH were measured in situ (Table B1).

Physicochemical data on the discharged effluent from the wastewater treatment plant of the city of Ioannina were collected from http://astikalimata.ypeka.gr/ and were averaged for a period of ten days prior of the sampling dates (Table B2). Water flow rates through the hydroelectric dam were obtained from the local managing company (Public Power Corporation, Department of Renewable Energy). Flow rates were available only for periods when the dam doors were open (December-March) and varied from 236.51 to 808.91 $m^3 s^{-1}$, exceeding the mean annual flow of the river as previously described. Thus, we used the number of days that the dam was closed before each sample was taken (factor Day), as a proxy for the river flow (i.e. in February that dam was open during sampling Day=0).

3.3.2 DNA extraction and sequencing

DNA extraction was performed with the MoBio Power Soil kit. Tag-pyrosequencing of the V1-V3 region of the 16S rRNA gene was performed with the primer pair 27F (5'- AGRGTTTGATCMTGGCTCAG-3') and 519R (5'-GTNTTACNGCGGCKGCTG-3') for Bacteria [22]. For whole genome sequencing, 5 µg of a DNA were used with Nextera XT and sequenced in an Illumina HiSeq 2000 (150 bp single end reads), according to established protocols by the manufacturer.–Trimming, quality control and further processing of the 454 sequences was performed with the MOTHUR software (v1.30) [23], and classification using the SILVA SSU database (v119) [24]. Operational Taxonomic Units (OTUs) were built by hierarchical clustering (average neighbor algorithm) at 97% sequence similarity. Coverage values were calculated with Good's formula using MOTHUR. Sequences have been deposited in the NCBI Short Read Archive (SRA) under accession number SRS652401.

3.3.3 Quality trimming and metagenomic assembly

Illumina reads were trimmed with Q=15 Phred using SolexaQA [25] and only reads longer than 50pb were considered for further analysis. The coverage of the metagenomic datasets was estimated using Nonpareil [26]. Assemblies were performed using a hybrid protocol as previously described [27]. Protein-coding genes were predicted using MetaGeneMark.hmm [28]. Sequencing and assembly statistics for the metagenomic datasets are provided in Table B1. Metagenomic 16S rRNA sequences were identified with Parallel-META (v.2.1) [29], and were subsequently processed for OTU picking (>97% identity threshold) and taxonomic identification with SILVA database [24] using MacQIIME 1.8.0 using default parameters [30]. The MyTaxa algorithm was used with default parameters [31] to determine the taxonomy of the archaeal and bacterial contigs. Predicted genes were functionally annotated based on their best match against the SEED database using the subsystems categories [32]. Only Blastp [33] searches with minimum 50% identity, 50% query sequence coverage, and an e-value of 10^{-4} cut-offs were used.

3.3.4 OTUs assignment to different habitats

All sequences from the SILVA 119 database, with available information on the source of origin were used to build a database of representative sequences with a habitat assignment. Each reference SILVA sequence was classified in house into seven different categories, based on the associated project: *Marine* (combining coastal and open ocean sequences, marine sediments, marine animals), *Freshwater* (lakes, rivers, streams),*Aquatic* (estuaries, ground water, various aquatic environments) *Terrestrial* (soil, plant, and animal-associated sequences), *Human gut* and SSW (containing sequences from sludge, sewage and wastewater samples) and *Others* for all other habitats, or when no habitat information was available. The SILVA sequences were subsequently clustered with cd-hit in 97% identity clusters, aiming to represent species level OTUs, and each cluster was assigned to a habitat based on the consensus habitat classification of the sequence(s) it contained. For example, a 97% identity cluster assigned to the habitat *'Freshwater'* contained 16S rRNA sequences that were all and exclusively found in freshwater habitats, representing a typical freshwater species. Clusters containing sequences from different habitats were assigned to the category *Others*. The sequence representatives from each cluster were used as a reference database, and each OTU rRNA sequence from Kalamas was then assigned to a habitat based on a BLAT search and a >99% nucleotide sequence identity cut-off. In order to evaluate the validity of our database a similar analysis was performed with 454 amplicon 16S rRNA sequences from seawater surface samples from the Gulf of Mexico [34] and soil samples from the Kessler Farm Field Laboratory [35]. Reference database is available from the authors upon request.

3.3.5 Determining differentially abundant features

Metagenomic reads were mapped on the predicted genes from the assembly using BLAT [36] with at least 95% identity and 50% of query length aligned. The abundance of each gene on each dataset was estimated by the number of reads that mapped on the genes with the above cutoffs (gene coverage). The relative abundance of annotation terms (subsystems) or taxa (phylum, class and genus level) in each dataset was estimated based on the sum of the abundances of the corresponding genes,

normalized for dataset size (*i.e.*, total number of reads) using the DESeq2 package version 3.0.2 [37]. Differentially abundant categories (taxa or subsystems) between samples were identified with DESeq2 using the binomial test and false discovery rate <0.05.

3.3.6 Metagenomic comparison of aquatic datasets

Metagenomic datasets from Lake Lanier [38], Amazon River [8], and Chesapeake and Delaware bay [39] were trimmed with PRINSEQ (http://edwards.sdsu.edu) and were assembled similarly with the Kalamas river metagenomes. Comparisons between datasets were performed based on the gene counts of the different functional annotation terms (subsystems). To make these counts comparable between datasets, protein sequences were clustered using the CD-HIT algorithm [40] with the following parameters: S=97 (similarity threshold) and aL=0.5 (minimum length coverage). Representative proteins from each cluster were annotated based on their best match against the SEED database. Comparisons between differentially abundant subsystems were performed using DESeq2 as described above. Principal Coordinates Analysis (PCoA) on the median-normalized counts (by DESeq2) of subsystems was performed using the function cmdscale in R. Searches for human gut associated species were performed based on comparisons of predicted protein sequences against the Integrated Gene Catalog (IGC) of genes related with the human gut microbiome [41]. A gene match against IGC was defined as at least 95% cutoff identity and 50% query length coverage.

3.3.7 Diversity indices and multivariate analysis

Diversity indices were calculated for the 97% identity OTUs constructed from the 16S rRNA gene amplicon sequences, or the 16S rRNA-encoding short reads recovered from the metagenomes. Coverage values were calculated using MOTHUR (v1.30) [23]. Shannon and Simpson (1-D) diversity indices, Dominance (D) and Richness of observed OTUs were calculated using PAST [42]. Non-metric Multidimensional Scaling (NMDS) analysis and Cluster analysis were performed as described in [43]. The significance of environmental parameters for the ordination of the samples was calculated using the function envfit of the R package vegan.

3.4 RESULTS

3.4.1 Bacterial community structure of Kalamas river

A total of 2,929 Operational Taxonomic Units (OTUs, 97% identity clusters) were identified from all the 454 datasets, containing 96,500 rRNA sequences in total (Table 3.1). Only 24% of these OTUs had a >99% identity match against the SILVA database, while the rest represented 'novel' OTUs (Table 3.1). Those novel OTUs comprised a significant portion of the total community, accounting for 13%-50% of the total reads in most of the samples.

Table 3.1: Taxonomic diversity of Kalamas samples.

Diversity indexes were estimated using the 16s rRNA amplicons (top table) or the short 16S rRNA sequences that were identified from the WGS metagenomic datasets. In both cases OTUs were constructed as 97% sequence identity clusters. * [95% Conf. Interval].

	Diversity Index	kal1-nov	kal2-nov	kal3-nov	kal1-feb	kal2-feb	kal-3feb	kal1-may	kal2-may	kal3-may
16s rRNA amplicon sequences	# OTUs	611	480	364	677	832	945	471	220	312
	Chao1	625.9	526.9	404	729.4	885.6	989.8	484.1	237.3	328.33
	Chao1 lower boundary*	618.5	618.5	618.5	618.5	618.5	618.5	618.5	618.5	618.52
	Chao1 up. boundary*	640.8	640.7	640.7	640.7	640.7	640.7	640.7	640.7	640.78
	% total reads of classified OTUs	30,4	33,13	39,29	31,17	31,13	31,11	84,08	43,18	45,51
	# Sequences	8598	12544	5731	7032	7771	8873	15098	13,15	17,695
	Dominance_D	0.05	0.091	0.038	0.038	0.028	0.033	0.088	0.117	0.209
	Simpson_1-D	0.95	0.909	0.962	0.962	0.97	0.967	0.912	0.883	0.791
	Shannon_H	4.67	3.49	4.22	4.76	5.06	5.05	3.70	2.859	2.405
	Coverage	0.96	0.97	0.96	0.95	0.94	0.93	0.98	0.99	0.98
Metagenomes	OTUs	2632	2431	3006	1353	1730	3010	2677	1480	1795
	# Sequences	19203	32036	30002	3970	5448	13927	60160	26781	26186
	Dominance_D	0.007	0.008	0.005	0.003	0.003	0.003	0.021	0.044	0.016
	Simpson_1-D	0.993	0.992	0.995	0.997	0.997	0.997	0.979	0.956	0.984
	Shannon_H	6.635	6.003	6.515	6.511	6.698	7.030	5.512	4.700	5.448
	Coverage	0.96	0.97	0.97	0.81	0.83	0.91	0.98	0.97	0.97

Both pyrosequencing and WGS datasets revealed the dominance of bacterial sequences, with only a 0.03 % to the total reads having archaeal origin. Higher diversity as indicated from observed OTU richness, Chao and Shannon indices, was observed during February, while the lowest diversity was observed in May at the dam (kal2) and the estuary (kal3) sites (Table 3.1). Good's coverage values exceeded 0.95 for both methods in all sites apart from the samples collected during February, indicating that approximately 95% of the total 16S OTUs were recovered by sequencing. The higher diversity during February was also reflected in assembly of the WGS datasets (Table B1): only 20% of the short metagenomic reads assembled in contigs longer than 500bp for the February metagenomes, while about 50% of the total reads assembled for the May samples. The coverage of the microbial community present in each sample achieved by the corresponding metagenomic dataset was estimated based on the redundancy of the reads using the Nonpareil method [26] and confirmed the higher complexity of the February samples (coverage=0.36) as opposed to the other months (0.8 for kal2nov) (Table B1).

Cluster analysis performed on the abundance distributions of the identified OTUs, revealed that the February samples had the lowest spatial differences exhibiting Morisita similarities above 84% (Fig. 3.1B). The highest spatial differences were observed during May between the estuary (kal3) and the other sites, and in November between the exit of the ditch (kal1) and the rest of the sites. Overall, temporal differences were more pronounced than spatial ones, since Morisita similarities, at temporal level, were lower than 50% in all sites (Fig. 3.1B). While temporal differences exceeded spatial ones, all samples showed extensive variability across sites and months as revealed when comparing the number of OTUs that were shared among samples (Fig. 3.2). From the total 2,929 identified OTUs only two were detected in all 9 samples, and ten OTUs were detected in at least 7, comprising less that 3% of the total reads per sample. When comparing all three sites during the same month, 61, 229 and 40 OTUs were detected in all three sites during November, February and March respectively. Given the similar sequencing depth across samples (Table 3.1), the comparison of the number of detected shared OTUs can provide a picture of the size and establishment of a "core" bacterial community in the river. Based on the available Kalamas datasets, the majority of the identified OTUs were comprising local communities, with a very small fraction

reoccurring through the months, or being detected across the river, with the possible exception of February samples (Fig. 3.2).

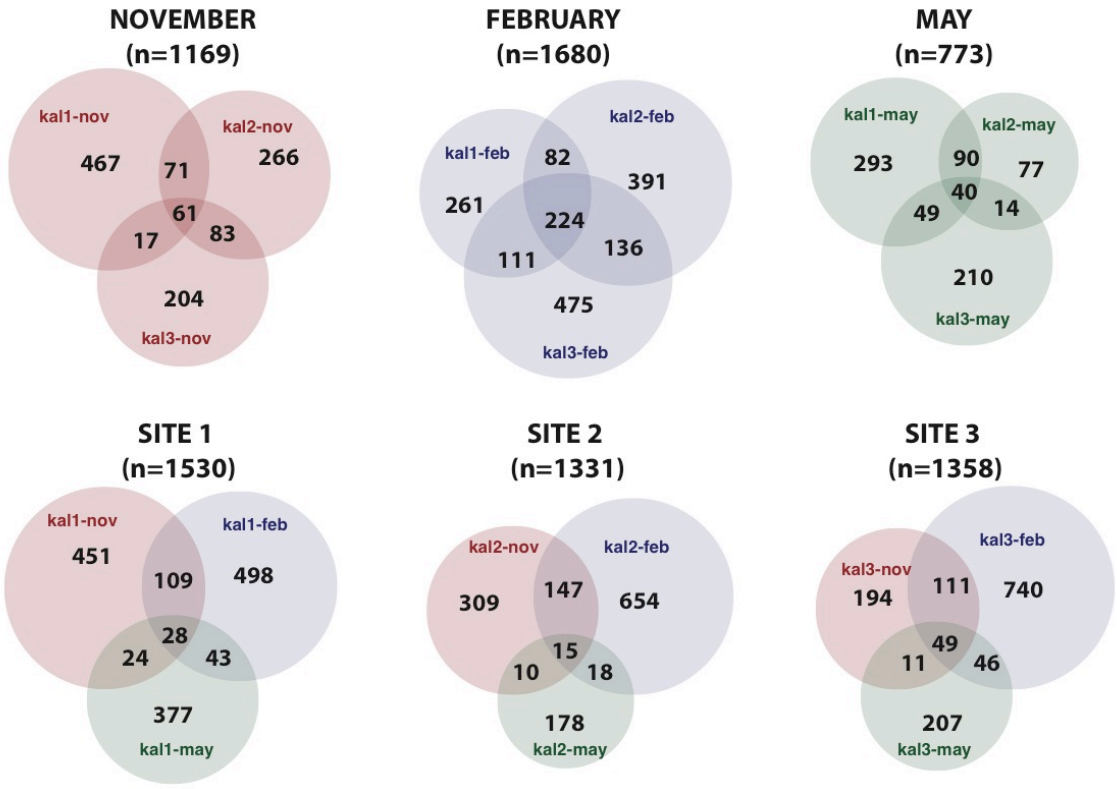


Figure 3.2: Re-occurrence of detected bacterial OTUs through time and space. Venn diagrams showing the sets of detected bacterial OTUs in each sample, and their intersections. The top diagrams represent comparisons among the three different sites, for each month, and the bottom diagrams represent the temporal comparisons for each site. Only 2 OTUs out of the total 2929 were detected within the intersection among all nine samples. Notice that only during February a common bacterial community is detected across sites (224 OTUs), comprising ~30 and up to 50% of the total OTU diversity found in each of the sites. During November and May, site specific OTUs comprise the majority of the bacterial diversity and minimum overlap is observed. Similarly, only a small fraction of the community reoccurs during the different months in the same sites (bottom diagrams).

Taxonomic composition was estimated based on (i) the bacterial 16S rRNA gene amplicon sequences, (ii) the short reads encoding 16S rRNA gene fragments recovered in the metagenomes, (iii) the taxonomic classification and abundance of predicted protein-coding genes from metagenomes, and where consistent among all three methods. *Proteobacteria* and *Bacteroidetes* were the most abundant phyla across all samples. Class level taxonomic distributions revealed the high abundance of *Betaproteobacteria* (>20% of sequences) in all samples during November and February, followed by members of *Flavobacteria* (>10%) (Fig. 3.1C). During May, *Gammaproteobacteria* dominated in kal1 and kal2 (40%), followed by *Betaproteobacteria*, while *Bacilli* was the dominant Class in the estuary samples (Fig. 3.1C). *Actinobacteria* were observed at higher abundances only during February, ranging from ~20% in the 16S rRNA amplicons dataset to ~10% in the metagenomic dataset (Fig. 3.1C).

3.4.2 Assignment of potential source 454 amplicons

In order to quantify the relative abundance of autochthonous vs. allochthonous bacteria in the riverine samples, we constructed an in house database containing all available SILVA119 16S rRNA reference sequences with associated metadata [24], and assigned each SILVA sequence to a specific a habitat. The SILVA sequences were subsequently clustered in 97% identity clusters, aiming to represent species level Operational Taxonomic Units (OTUs), and each cluster was assigned to a habitat based on the consensus habitat classification of the sequence(s) it contained (see experimental procedures). Subsequently, the Kalamas OTUs were assigned to a likely source habitat based on high nucleotide identity (99%) blastn matches against the constructed in house reference database.

Two control samples were first evaluated in order to test the robustness of the database and our approach for assigning sequences to habitats: a typical open ocean sample from the Gulf of Mexico [34] and a soil sample from the Kessler Farm Field Laboratory [35]. About 67% of the sequences from the ocean samples were classified in aquatic OTUs (mostly seawater, Fig. 3.3), while the other categories were almost

undetected. Similarly the control soil samples comprised of ~65% typical terrestrial OTUs, while the other categories were almost undetected.

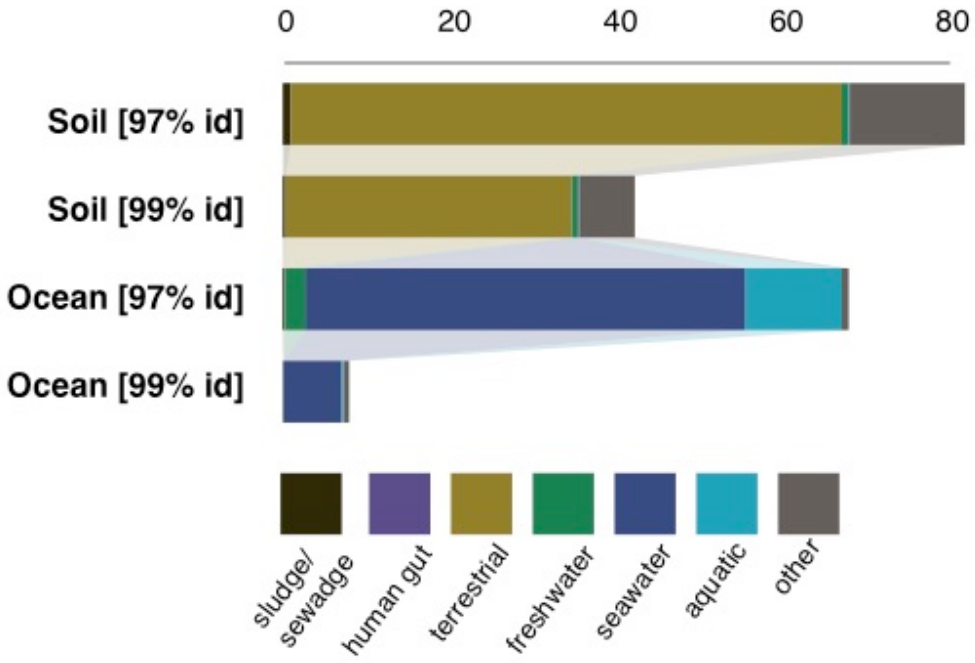


Figure 3.3: Habitat assignment of 16S rRNA gene sequences for control datasets. Seawater surface and soil samples were used for evaluation of the habitat assignment protocol, based on nucleotide similarities (97% and 99%) to reference sequences (SILVA119) associated to specific habitats. Notice that the majority of the soil sample is comprised of typical terrestrial OTUs, while the open ocean sample is comprised mostly of typical seawater or other aquatic OTUs. When using an identity cutoff of 97% the majority of the OTUs can be classified, since they have a match against the in house SILVA database. When the threshold is higher (99%) only a small fraction can be classified, nevertheless the distribution of habitats is very similar.

The habitat assignment analysis for the Kalamas 16S rRNA datasets revealed that typical freshwater OTUs (mostly *Actinobacteria*) were highly abundant during February, summing up to more than 40% of the total sequences (Fig. 3.1D). Freshwater-, and generally aquatic-related OTUs were also found in high abundance during

November in kal2 and kal3, but were almost absent in kal1, for which the majority of OTUs were assigned in the Sludge/Sewage/Wastewater (SSW) group (Fig. 3.1D), mostly classified as *Alphaproteobacteria* (genus *Ancalomicrobium*) and *Betaproteobacteria* (Family *Oxalobacteraceae*). Human gut associated OTUs reached their maximum in kal1nov, comprising 0.12% of the total detected sequences, decreased downstream and were almost undetected in the rest of the samples. Decreased abundance of typical freshwater taxa was also observed during May, with the upstream samples showing predominance of bacteria typically encountered in terrestrial environments, with the most abundant OTUs in kal1/2 classifying to *Gammaproteobacteria* (genus *Acinetobacter*). Interestingly, the abundance of terrestrial, human gut or SSW-related OTUs generally decreased downstream in all months, with the exception of February, during which all three sites had very similar distributions and freshwater taxa were dominant. Collectively, these results revealed that the influence of municipal sewage can be detected even at 10s to 100s of kilometers downstream its presumptive source (the Lapsista ditch), and that, at least temporarily, the river community could be dominated by non-typical aquatic organisms, indicating the prevalence of allochthonous inputs.

3.4.3 Microbial functional diversity in Kalamas river

Functional annotation against the SEED sybsystems resulted in the classification of 16.6% (kal1feb) to 40.99% (kal3nov) of total reads (Table B1). Pairwise comparisons of the functional distributions from all samples revealed relatively similar gene content with only 171 out of 1,737 identified subsystems exhibiting significantly ($p<0.05$) differential abundances between any two samples. Grouping of the 171 differentially abundant subsystems into broader functional categories revealed several notable shifts in functions over time and space (Fig. 3.4A).

Most of the differences in functional gene content were observed between samples from different months, in accordance with seasonal variation in the taxonomic distributions (Fig. 3.4B). Nitrogen metabolism genes were more abundant in kal1 during November while these genes remained at very low levels during the other months. February samples overall exhibited significantly higher abundances of autotrophy and

DNA replication related genes, and the lowest abundances of virulence genes (Fig. 3.4A). During May the highest levels of virulence genes were observed. Genes for polyhydroxyboutyrate metabolism (short chain fatty acids transporter, D-beta-hydroxybutyrate permease) and respiration genes were also enriched during May and in kal1nov. Genes for NiFe hydrogenases, implying anaerobic conditions, were mostly enriched during November, while genes related to terminal cytochrome oxidases were prevalent in May (Table B3).

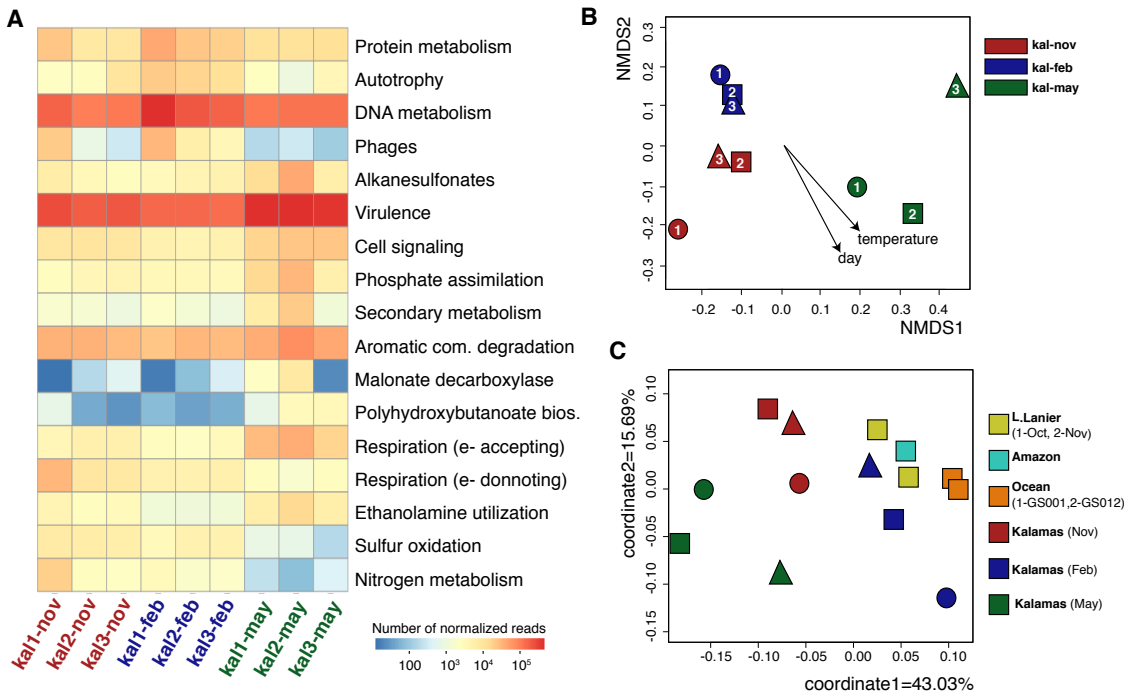


Figure 3.4: Functional profiles of Kalamas River microbial communities.

(A) Subsystems (SEED database) with statistically significant differences in abundance (Negative binomial test, DeSEQ) through space or time. (B) NMDS plot of functional distributions based on normalized abundances of genes classified at different SEED subsystems, and of taxonomic distributions based on abundances of OTUs (97% similarity) constructed from the metagenomes rRNA 16S gene-encoding reads. Arrows indicate significance ($p<0.02$) for the plotting of the samples (C) Principal coordinates analysis of Kalamas and other freshwater and estuarine samples based on gene content and taxonomic distributions. The gene content distributions were derived from the number of different genes that could be assigned to a subsystem. The taxonomic

distributions reflect the abundances of 97% OTUs, constructed from the 16S rRNA metagenomic reads.

Overall the functional differences across sites and months reflected the taxonomic differences, indicating that the taxonomic composition shifts are linked to functional gene content differences. Non-metric multidimensional scaling (NMDS) of the functional distributions revealed that temperature and the factor Day (number of Days that the Dam was closed before sampling) were the most significant factors ($p<0.02$) for the ordination of the samples, while the distance from the exit of the ditch showed no significance (Fig. 3.4B). Similar findings were noted in NMDS analysis of taxonomic distributions (Fig. 3.4B). The significance of temperature (indirectly related with season) was consistent with the predominance of temporal over spatial variation. The significance of Day (a proxy for the river flow) reflected the effect of river flow on functional and taxonomic distributions.

3.4.4 Comparison of taxonomic and functional diversity with other ecosystems

Principal Coordinates Analysis (PCoA) of the subsystem-based functional distributions (Fig. 3.4C) further confirmed the seasonal separation of the Kalamas samples and revealed a higher similarity of the February samples with samples from the Amazon river and the typical temperate Lake Lanier, compared to November and May samples (Fig. 3.4C). The Chesapeake and Delaware estuary samples were similar between them but were highly differentiated from all Kalamas estuary samples (kal3).

More specifically, 125 subsystems exhibited significant ($p<0.05$) differences in genes abundance between Kalamas samples and those of other habitats. All the subsystems that were characteristic for the May upstream samples compared to the other Kalamas samples (metabolism of aromatic compounds, virulence, malonate decarboxylase, stress response) were similarly characteristic of Kalamas when May samples were compared to the other freshwater habitats. Genes related to autotrophy (CO dehydrogenases, chlorophyll biosynthesis, Photosystems I and II) were less abundant in Kalamas samples, which were instead characterized by more heterotrophic

metabolism functions. Only the February samples, which were more similar to the samples from Lake Lanier (Atlanta, USA), had more autotrophic signatures. In accordance, the taxonomic composition of the Kalamas February samples had the highest similarity with the Lake Lanier ones (Fig. 3.4C), mostly attributed to the increased relative abundances of representatives of the hgcI clade of *Actinobacteria* and the decreased abundances of *Gammaproteobacteria* relative to other Kalamas samples. Members of *Bacteroidetes* were dominant in most Kalamas samples, while they were detected in lower abundances in Amazon and Lake Lanier samples.

Some of the Kalamas samples also exhibited high abundance of microbial human gut (HG) signals, compared to the other ecosystems. The HG genes reached a maximum abundance of 0.17% (of total genes) in November, close to the ditch, and decreased downstream, only slightly increased from kal1 to kal2 during May and were almost undetectable in the estuary, and in all February samples (Fig. 3.5). The majority of the detected HG associated genes were affiliated with *Gammaproteobacteria*, the families of *Enterobacteriaceae, Aeromonadaceae, Vibrionaceae,* and *Shewanellaceae,* while in terms of functions, the majority of them were annotated as hypothetical. These results were also, in general, consistent with the 16S rRNA gene-based habitat assignment reported above.

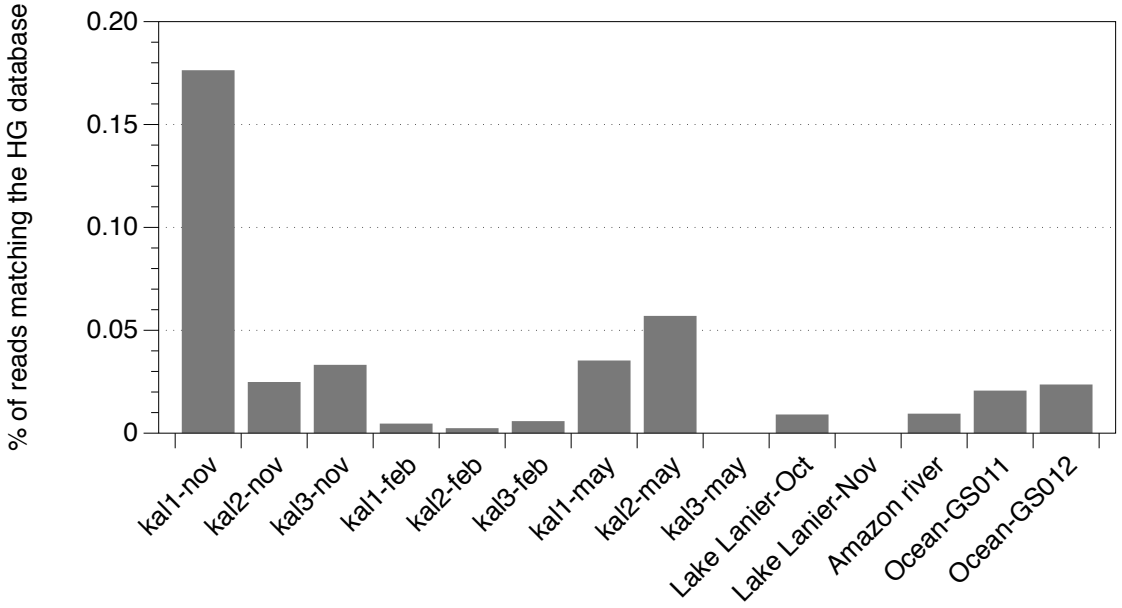


Figure 3.5: Quantification of HG associated sequences in aquatic habitats.
The graph shows the relative abundance of genes that had >95% nucleotide identity blastn matches against blastp match against the human gut database.

3.5 DISCUSSION

In this study we investigated the spatial and temporal changes in bacterial taxonomic and functional diversity along a river to estuary gradient. Seasonal taxonomic diversity differences were more pronounced than spatial ones (Fig. 3.1). Functional distributions were in general consistent with the taxonomic ones and they were both mostly attributed to temperature and the function of the dam (Fig. 3.4B), reflecting the effects of season and river flow, respectively.

During the three months examined, Kalamas river showed a prevalence of a typical freshwater bacterial community only during February: the taxonomic diversity was estimated at its highest, and the majority of OTUs were identified as typical freshwater

organisms (Fig. 3.1D). This findings agreed with a previous study in Danube River, which employed a similar approach to the one described here for habitat assignment and reported that more than 40% of the bacterial population was assigned to typical freshwater taxa [5]. However, spatial patterns observed in other riverine ecosystems, where freshwater related taxa increased in abundance downstream [5,12], were not observed in Kalamas river in any of the seasons sampled, presumably due to the different physical properties (e.g., length, depth) and flow rate of Kalamas, and the (more pronounced) effects of allochthonous inputs (see also below).

While *Betaproteobacteria* were observed in almost all samples (Fig. 3.1C), corroborating the findings of all recent studies of riverine ecosystems [5,8,12,13,44,45], *Actinobacteria*, a major group in riverine ecosystems, were abundant in Kalamas River only during February (Fig. 3.1C). In accordance with the high freshwater signal, February samples resembled those of a typical temperate lake (L. Lanier) and the relative pristine sample from the Amazon River in terms of gene content, with comparable abundances of genes related to autotrophy (Chlorophyll biosynthesis, photosystems I and II and CO_2 uptake) and DNA metabolism. Autotrophy related genes, were primarily (>90%) assigned to *Cyanobacteria*, while DNA replication genes were mostly assigned to *Actinobacteria*, implying that the prevalence of the previous gene functions during February and the high similarity with other freshwater habitats could be largely attributed to these two typical freshwater groups that were almost absent in the other months (Fig. 3.1C). Furthermore, the February samples exhibited the highest similarity across sites, and even the estuary sample exhibited prevalence of typical freshwater organisms, as opposed to seawater (Fig. 3.1D). The dominance and establishment of a typical freshwater community throughout the river during February could be attributed to the high water flow during this month (261.28 m^3 s^{-1}), which was significantly higher than the mean annual flow of the river (54 m^3 s^{-1}). During February sampling, the dam was open, and the precipitation was at its highest (Fig. 3.6), conditions leading to high water flow and resulting in the high homogeneity of the three sites. Indeed only during February, a core community of bacterial OTUs could be detected throughout the river, even at the estuary samples, which exhibited noticeably lower salinity compared to other months (Table B1).

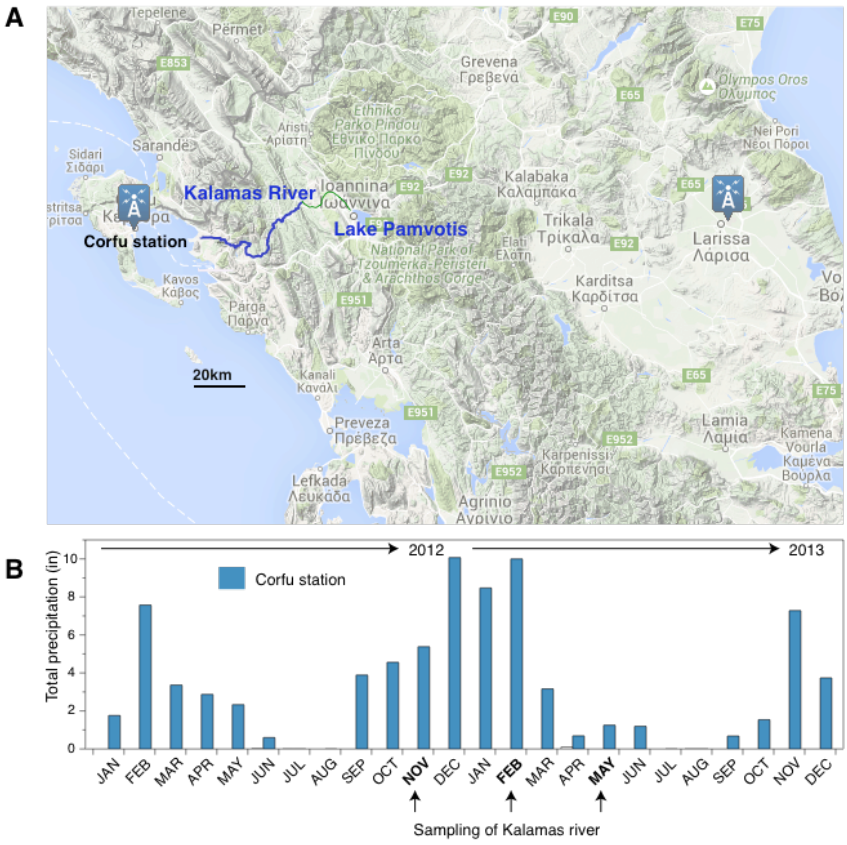


Figure 3.6: Precipitation data during the sampling expedition.

A. Location of the nearest to the Kalamas river stations from which precipitation data were acquired. B. Monthly total precipitation during 2012 and 2013. Data were acquired from the NOAA/ESRL Physical Sciences Division, Boulder Colorado from their Web site at http://www.esrl.noaa.gov/psd/.

In contrast to the February observations, the river showed high spatial variation and lower diversity during November and May. The upstream sites showed a dominant signal of Sewage-Sludge-Wastewater (SSW) and terrestrial OTUs, during November and May respectively, while increased numbers of seawater OTUs were detected at the estuary (Fig. 3.1D). A notable increase in autotrophy related genes in November and May close to the estuary (Fig. 3.4A) was consistent with the previous observations of enrichment of autotrophy genes in marine compared to freshwater samples [8]. Overall the functional and taxonomic diversity in Kalamas estuary was dissimilar to other estuarine

samples (Fig. 3.4C), but any conclusions from this comparison should be interpreted with caution, since Chesapeake and Delaware estuaries are formed from larger and more eutrophic rivers than Kalamas River.

Increased abundances of Sewage-Sludge-Wastewater (SSW) and terrestrial OTUs (Fig. 3.1D) as well human gut genes (Fig. 3.5) close to the ditch (kal1) in November and in kal1 and kal2 in May suggested a seasonally variable impact of allochthonous inputs. This observation was in accordance with the dominance of aerobic and anaerobic respiration genes and diverse pathways related to complex organic compounds degradation (Fig. 3.4A).–Terrestrial, agricultural and industrial inputs in Kalamas could also underlie the high abundance of additional secondary metabolism pathways such as malonate decarboxylase [46,47] and alkanesulfonate assimilation and utilization genes [48] in kal2may, and presumably supported the presence of the highly heterotrophic bacterial communities observed. The dominance of *Acinetobacter* related OTUs and genes during May, especially in kal2may, combined with the low flow rates and precipitation during this month (Fig. 3.6) and the low diversity indices observed (Table 3.1) imply 'species sorting' as the mechanism that may have fostered the high prevalence of this taxon, which is most likely of terrestrial origin based on 16S RNA gene-based habitat assignment and metagenomic data. This finding indicates the potentially significant impact of allochthonous species in the structure of water bacterial communities in rivers, constituting efforts trying to delimit river bacterioplankton rather challenging [49,50].

Increased values of BOD5, COD, Total Nitrogen and NH4 at the effluent of the wastewater treatment plant of Ioannina during November, were accompanied by the highest signal of Human Gut Database (HG) sequences observed in kal1nov (Fig. 3.5), high numbers of nitrogen metabolism and virulence related genes (Fig. 3.4A), as well as SSW related OTUs (Fig. 3.1D). Detection of HG associated genes in freshwater environments can be potentially used to quantify the impact of sewage inputs in streams. Thus, the prevalence of HG and SSW related OTUs, could serve as biomarkers of the anthropogenic impact on the riverine taxonomic and functional diversity in this sample. The decrease of the HG signal downstream and the abundance of HG genes in kal1 relative to other freshwater ecosystems (Fig. 3.5) supported the potential origin of HG

genes from the city of Ioannina and their transport to Kalamas River through the Lapsista ditch. Dilution of HG signal along the river in November could be explained by competition with other bacteria better adapted to freshwater environment, and it is consistent with an increase of aquatic related OTUs downstream.

Comparisons of the detected OTUs across all samples revealed an extensive variability, and no core microbial community could be established among all sites and months (Fig. 3.2). Only two out of thousands of identified OTUs could be detected across sites and samples, revealing the highly dynamic nature of the river. In contrast, previous seasonal studies on typical freshwater ecosystems like the Lake Lanier had shown relatively stable gene contents over time and after big summer storm events [51]. In Kalamas River and generally for streams and mid-size rivers, the extensive variability is not unexpected, since variable hydrology factors and temporally unpredicted 'mass effects' can influence cell dispersal [4,52] and thus diversity and function. The occurrence of a core bacterial community has been previously reported in high volume streams such as the Danube rivers [5] and Thames [12]: high retention times can potentially allow the establishment of competitive freshwater taxa in the latter ecosystem. Savio and colleagues reported the detection of an increasingly abundant core bacterioplankton community along the 2,600 km length Danube, and hypothesized that it is due to the decrease of 'mass effects' from the riparian zone (allocthonus inputs). Read and colleagues reported increased abundances of *Bacteroidetes* upstream, which are gradually being replaced with the more typical freshwater *Actinobacteria* downstream, attributing this pattern to ecological succesion [12]. Similar patterns were not observed in Kalamas River, possibly due to the smaller catchment area that may intermittantly increase 'mass effects' from the land or from the Lapsista ditch. Additionally the high variability in the river flow attributed to the function of the dam and/or local climatic factors (intense precipitation differences between wet and dry season, Fig. 3.6) that can influence retention times and potentially 'species sorting'.

The differentiation of Kalamas taxonomic diversity from other riverine ecosystems was apparently attributable to the increased allochthonous inputs influencing bacterial diversity in Kalamas River. Thus, efforts to investigate Kalamas bacterial diversity in the broader context of river bacterioplankton biogeography should

take into account these allochthonous inputs. Results regarding taxonomic diversity fluctuations resemble those of Portillo and colleagues [9], showing that intermittent changes of biogeochemical conditions drive taxonomic diversity in streams, without specific seasonal patterns. Similarly, in our study, habitat assignment of the identified OTUs, functional diversity and comparisons with other datasets and the human gut database showed that bacterial communities along a small temperate river can be strongly influenced by external inputs, which may vary across seasons or sites, overriding the presence of typical freshwater OTUs or a 'core' bacterial community that was undetectable in this study. Additionally hydrological factors such as river flow also affected significantly taxonomic and functional bacterial diversity, either by assisting the establishment of local communities (species sorting) under conditions of low flow, enhancing the dominance of allochthonous species, or by preventing the establishment of allochthonous species in conditions of high flows. In summary, our study further confirmed the dynamic character of riverine ecosystems, the significance of human activities in shaping bacterial diversity in rivers and streams, and highlighted the need for more detailed and in depth studies in order to better monitor and quantify the influence and the fate of external anthropogenic inputs.

3.6 REFERENCES

1. Shiklomanov, I. A. & Rodda, J. C. *World Water Resources at the Beginning of the Twenty-First Century*. (Cambridge University Press, 2004).
2. Cole, J. J. *et al.* Plumbing the Global Carbon Cycle: Integrating Inland Waters into the Terrestrial Carbon Budget. *Ecosystems* **10,** 172–185 (2007).
3. Battin, T. J. *et al.* The boundless carbon cycle. *Nat. Geosci.* **2,** 598–600 (2009).
4. Findlay, S. Stream microbial ecology. *J. North Am. Benthol. Soc.* **29,** 170–181 (2010).
5. Savio, D. *et al.* Bacterial diversity along a 2600 km river continuum. *Environ. Microbiol.* **17,** 4994–5007 (2015).
6. Fortunato, C. S. *et al.* Determining indicator taxa across spatial and seasonal gradients in the Columbia River coastal margin. *ISME J.* **7,** 1899–1911 (2013).
7. Winter, C., Hein, T., Kavka, G., Mach, R. L. & Farnleitner, A. H. Longitudinal Changes in the Bacterial Community Composition of the Danube River: a Whole-River Approach. *Appl. Environ. Microbiol.* **73,** 421–431 (2007).
8. Ghai, R. *et al.* Metagenomics of the Water Column in the Pristine Upper Course of the Amazon River. *PLOS ONE* **6,** e23785 (2011).
9. Portillo, M. C., Anderson, S. P. & Fierer, N. Temporal variability in the diversity and composition of stream bacterioplankton communities. *Environ. Microbiol.* **14,** 2417–2428 (2012).
10. Marti, E., Jofre, J. & Balcazar, J. L. Prevalence of Antibiotic Resistance Genes and Bacterial Community Composition in a River Influenced by a Wastewater Treatment Plant. *PLOS ONE* **8,** e78906 (2013).
11. Staley, C. *et al.* Bacterial community structure is indicative of chemical inputs in the Upper Mississippi River. *Front. Microbiol.* **5,** 524 (2014).
12. Read, D. S. *et al.* Catchment-scale biogeography of riverine bacterioplankton. *ISME J.* **9,** 516–526 (2015).
13. Staley, C. *et al.* Core functional traits of bacterial communities in the Upper Mississippi River show limited variation in response to land cover. *Front. Microbiol.* **5,** (2014).
14. Chao, Y. *et al.* Metagenomic analysis reveals significant changes of microbial compositions and protective functions during drinking water treatment. *Sci. Rep.* **3,** (2013).
15. Gillings, M. R. *et al.* Using the class 1 integron-integrase gene as a proxy for anthropogenic pollution. *ISME J.* **9,** 1269–1279 (2015).
16. Luo, C. *et al.* Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. *Proc. Natl. Acad. Sci.* **108,** 7200–7205 (2011).
17. Kotti, M. E., Vlessidis, A. G., Thanasoulias, N. C. & Evmiridis, N. P. Assessment of River Water Quality in Northwestern Greece. *Water Resour. Manag.* **19,** 77–94
18. Albanis, T. A., Pomonis, P. J. & Sdoukos, A. T. Seasonal fluctuations of organochlorine and triazines pesticides in the aquatic system of Ioannina basin (Greece). *Sci. Total Environ.* **58,** 243–253 (1986).

19. Lekka, E. *et al.* Assessment of the Water and Habitat Quality of a Mediterranean River (Kalamas, Epirus, Hellas), in Accordance with the EU Water Framework Directive. *Acta Hydrochim. Hydrobiol.* **32,** 175–188 (2004).
20. Kagalou, I., Leonardos, I., Anastasiadou, C. & Neofytou, C. The DPSIR Approach for an Integrated River Management Framework. A Preliminary Application on a Mediterranean Site (Kalamas River -NW Greece). *Water Resour. Manag.* **26,** 1677–1692 (2012).
21. Economou, V. *et al.* Prevalence, antimicrobial resistance and relation to indicator and pathogenic microorganisms of Salmonella enterica isolated from surface waters within an agricultural landscape. *Int. J. Hyg. Environ. Health* **216,** 435–444 (2013).
22. Muyzer, G., Teske, A., Wirsen, C. O. & Jannasch, H. W. Phylogenetic relationships ofThiomicrospira species and their identification in deep-sea hydrothermal vent samples by denaturing gradient gel electrophoresis of 16S rDNA fragments. *Arch. Microbiol.* **164,** 165–172
23. Schloss, P. D. *et al.* Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl. Environ. Microbiol.* **75,** 7537–7541 (2009).
24. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35,** 7188–7196 (2007).
25. Cox, M. P., Peterson, D. A. & Biggs, P. J. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* **11,** 485 (2010).
26. Rodriguez-R, L. M. & Konstantinidis, K. T. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* **30,** 629–635 (2014).
27. Luo, C., Tsementzi, D., Kyrpides, N. C. & Konstantinidis, K. T. Individual genome assembly from complex community short-read metagenomic datasets. *ISME J.* **6,** 898–901 (2012).
28. Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* **38,** e132 (2010).
29. Su, X., Xu, J. & Ning, K. Parallel-META: efficient metagenomic data analysis based on high-performance computation. *BMC Syst. Biol.* **6 Suppl 1,** S16 (2012).
30. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7,** 335–336 (2010).
31. Luo, C., Rodriguez-R, L. M. & Konstantinidis, K. T. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Res.* gku169 (2014). doi:10.1093/nar/gku169
32. Overbeek, R. *et al.* The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. *Nucleic Acids Res.* **33,** 5691–5702 (2005).
33. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
34. King, G. M., Smith, C. B., Tolar, B. & Hollibaugh, J. T. Analysis of composition and structure of coastal to mesopelagic bacterioplankton communities in the northern gulf of Mexico. *Front. Microbiol.* **3,** 438 (2012).
35. Zhou, J. *et al.* Reproducibility and quantitation of amplicon sequencing-based detection. *ISME J.* **5,** 1303–1313 (2011).
36. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
37. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

38. Tsementzi, D., Poretsky, R., Rodriguez-R, L. M., Luo, C. & Konstantinidis, K. T. Evaluation of metatranscriptomic protocols and application to the study of freshwater microbial communities. *Environ. Microbiol. Rep.* **6,** 640–655 (2014).
39. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* **5,** e77 (2007).
40. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinforma. Oxf. Engl.* **28,** 3150–3152 (2012).
41. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32,** 834–841 (2014).
42. Hammer, O., Harper, D. & Ryan, P. Past: paleontological statistics software package for education and data analysis. *Paleontol. Electron.* 1–9
43. Meziti, A., Kormas, K. A., Moustaka-Gouni, M. & Karayanni, H. Spatially uniform but temporally variable bacterioplankton in a semi-enclosed coastal area. *Syst. Appl. Microbiol.* **38,** 358–367 (2015).
44. Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D. & Bertilsson, S. A Guide to the Natural History of Freshwater Lake Bacteria. *Microbiol. Mol. Biol. Rev.* **75,** 14–49 (2011).
45. Liu, Z., Huang, S., Sun, G., Xu, Z. & Xu, M. Phylogenetic diversity, composition and distribution of bacterioplankton community in the Dongjiang River, China. *FEMS Microbiol. Ecol.* **80,** 30–44 (2012).
46. Dimroth, P. & Hilbi, H. Enzymic and genetic basis for bacterial growth on malonate. *Mol. Microbiol.* **25,** 3–10 (1997).
47. Van Rossum, T. *et al.* Year-Long Metagenomic Study of River Microbiomes Across Land Use and Water Quality. *Aquat. Microbiol.* 1405 (2015). doi:10.3389/fmicb.2015.01405
48. Reichenbecher, W. & Murrell, J. C. Linear alkanesulfonates as carbon and energy sources for gram-positive and gram-negative bacteria. *Arch. Microbiol.* **171,** 430–438 (1999).
49. Leibold, M. A. *et al.* The metacommunity concept: a framework for multi-scale community ecology. *Ecol. Lett.* **7,** 601–613 (2004).
50. Satinsky, B. M. *et al.* Metagenomic and metatranscriptomic inventories of the lower Amazon River, May 2011. *Microbiome* **3,** 39 (2015).
51. Oh, S. *et al.* Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Appl. Environ. Microbiol.* **77,** 6000–6011 (2011).
52. Zeglin, L. H. Stream microbial diversity in response to environmental changes: review and synthesis of existing research. *Aquat. Microbiol.* 454 (2015). doi:10.3389/fmicb.2015.00454

CHAPTER 4

QUANTIFICATION OF INTRA-SPECIES GENE CONTENT DIVERSITY IN NATURAL BACTERIAL POPULATIONS

Reproduced in part with permission from D. Tsementzi[*], L. M Rodriguez-R[*] and K. T. Konstantinidis. A new method for the quantification of intra-population genetic diversity from metagenomes. In preparation. *Equal contribution authors.

4.1 ABSTRACT

Sequencing of hundreds of strains of the same species has allowed estimation of the species pangenome and shown that any two strains can frequently share as little as 60% of their total genes, likely reflecting adaptations to the environments from where they have been isolated (*e.g.*, human vs. environmental sources). To what extend this level of intra-species gene content variation is maintained within natural populations remains unclear but represents important information for better modeling microbial communities and the species concept. Here, we present a novel methodology for quantifying intra-population gene-content diversity based on modeling of the observed sequencing depth variations provided by a metagenome across the genome sequence or *de novo* genome assembly of a given population (metagenomic assembly). The method was validated based on simulations with >2,000 genomes and was applied to >1,000 populations from a variety of natural environments including soil, surface/deep ocean, lakes, human gut, and bioreactors. In general, natural populations were much more homogeneous than named species represented by isolate genomes (2 to 4 fold lower median gene-content diversity), although exceptions to this rule existed, and appeared to be taxon- as opposed to habitat-specific. In fact, the level of intra-population diversity appeared, on average, to be similar across habitats. Extensive gene content diversity was observed for ~20 aquatic populations, mostly *Bacteroides* and *Verrucucomicrobia*, far exceeding that of named species. The genetic mechanisms that

possibly underlie the latter patterns (e.g., lack of selective sweeps) as well as the relevance of the findings for the current species definition will be discussed.

4.2 INTRODUCTION

Bacterial strains belonging to the same species can often show extensive gene content variability, and the variable genes among strains can often encode drastically different phenotypes. For example, two *Escherichia coli* genomes can share as little as ~60% of their genes[1], with the rest 40% being variable and account for different phenotypical traits such causing severe disease in humans and animals vs. representing innocuous or commensal traits. Such extensive gene content variability within the same species is a common observation in sequence data from isolates[2–5], and has led to the definition of the "**pangenome**" concept[3,6,7], the collection of the total genes found in all genomes of a species. The pangenome includes the "core" genes found in all strains of the species, typically representing highly conserved metabolic functions that are essential for growth in all environments. Additionally, it includes the "flexible" or auxiliary genes, genes and operons that are present in some but not all the strains and are typically assumed to represent genetic adaptations to the different environmental conditions for which strains have been adapted to. For example, commensal *E. coli* strains contain flexible genes that determine pathogenicity phenotypes, which are not found in environmental *E. coli* strains[8].

The **size of the pangenome** and **proportion of flexible genes** greatly vary among different species, typically reflecting the ecological traits and evolutionary history of the organisms. For example, obligate pathogens such as *Bacillus anthracis* strains have nearly identical genomes with reduced gene content variability (Fig. 4.1), considered to be a recently emerged species with a narrow ecological niche[9]. Similarly, the human pathogen *Mycobacterium tuberculosis* is composed of nearly identical strains with low gene content and sequence variability[10,11], and those small differences in gene content among strains have phenotypic manifestations such as virulence and immunogenicity variations[12]. On the other hand, species distributed among more dynamic environments can exhibit large gene content and sequence variations among genomes, often times reflecting their adaptations to different niches. For example, the

genome collection of the opportunistic pathogen *Acinetobacter baumanni* isolates reveals high gene content variability among strains (Fig 4.1), consistent with the versatile lifestyle of the species and adaptation to both biotic and abiotic environments[13].

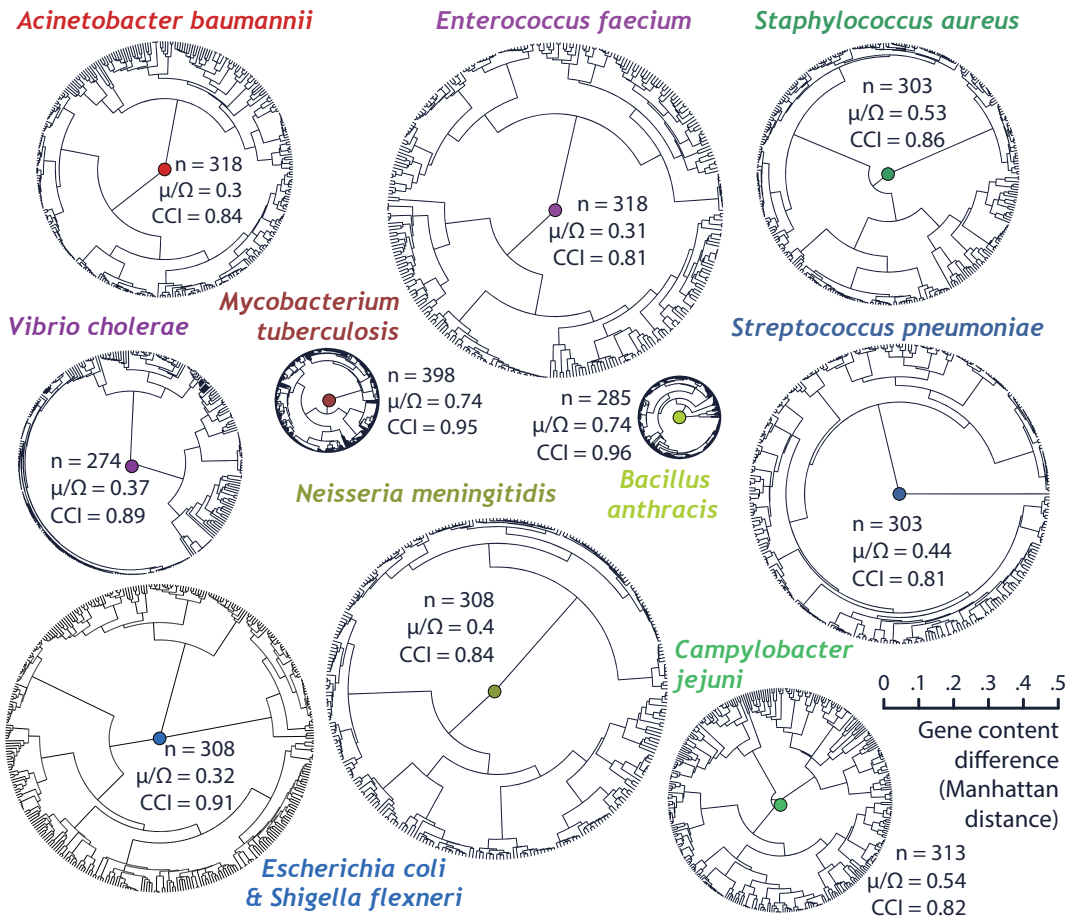


Figure 4.1: Intra-species gene content diversity in collections of bacterial isolates of named species.

Gene content diversity was measured in 10 species for which >300 reference complete genomes are available. Each tree represents the gene content dissimilarities between genomes of the same species expressed in Manhattan distances (Total number of non-core orthologhous genes (OGs) for each pairwise comparison over the total OG collection of the species). Thus the diameter of each tree is proportional to the maximum gene content dissimilarity between any two genomes. For each species the collective

gene content diversity estimates are given in two different metrics (see methods). μ/Ω=Average/Pangenome genome. CCI=Core Content Index. n=Number of genomes. Genomic and gene content statistics for the above genome collections are provided in Table 4.1. The trees were produced by hierarchical clustering of the Manhattan distances with the Ward method.

Traditionally, it was assumed that the differences in gene content among isolated strains of the same species are due to adaptations to the different environmental conditions from which the strains have been isolated (e.g., pathogenic *vs* commensal *E. coli* strains), thus bacterial strains of the same species that inhabit the **same environment should be more homogenous (clonal)** in their gene content. However, with the advent of high-throughput sequencing technologies and single cell genomes isolated from the same samples, it becomes apparent that extensive gene content variability might co-exist among cells of the same species within the same environment, collectively referred to as a natural bacterial **population**. For example, cells of the single *Prochlorococcus* species isolated from the same 1ml of water sample comprise hundreds of distinct subpopulations harboring unique flexible genes[14]. Thus, it remains unclear whether or not natural populations are more homogenous in their intra-population gene content relative to named species and by how much. Further, it remains also unclear how much of the diversity within natural populations arises from **adaptive *vs* neutral processes**, and what are the mechanisms that maintain it[15,16]? Characterizing the *in situ* diversity of bacterial populations has implications on the bacterial species concept; *i.e.*, what is the extend of population diversity that preserves species boundaries and how does this diversity change in response to environmental conditions[17–19]? In other words, if gene content variation is so extensive among cells of a single population, does our current bacterial species definition correspond to ecologically meaningful natural groups[20]?

Metagenomic based surveys provide the opportunity to assess the extent of within-population genomic diversity in a single sample and compare it across different environments. Provided adequate sequence coverage for a given population, intra-population diversity can be assessed by (a) *de novo* sequence assembly of abundant community members and/or (b) recruitment of metagenomic reads to available reference

genomes and assemblies. Application of those techniques to diverse systems has identified fine-scale genomic diversity within populations, both at the allelic diversity level (sequence identity variations) and at the gene content level[15,21]. For example, *Leptospirillum* strains of the same species with gene content variability have been identified through *de novo* metagenomic assemblies[15], and intra-population diversity has been linked to specializations to different micro-habitats within the acid main drainage system. Metagenomic read recruitments against reference genomes has revealed that hypervariable genomic regions, *i.e.*, genomic islands (regions present in the reference genome but absent from the metagenomic sample) are a common feature of natural populations[22], not only of isolates of named species. Such islands might be enriched in hypothetical proteins and represent neutral diversity in nature[19,23], or they might confer environmental adaptations[22]. The nature of the flexible genes in natural populations not only affects our interpretations of the species concept, but has also implications for understanding the role of this biodiversity in microbial community functioning and resilience. For example, metagenomic analysis of an enhanced biological phosphorus removal (EBPR) system captured the genetic diversity among coexisting *Accumulibacter phosphatis* populations that might partially explain the resilience and stable performance of the system in treatment applications [24]**.**

Therefore, quantification of intra-population gene content diversity *in situ* (*i.e.*, how much clonal or diverse the genomes of the same species are in the same environment) might elucidate underlying adaptative mechanisms, identify cases of neutral evolution and provide insights in the adaptability and resilience of microbial communities functioning. Here, we present a novel methodology for quantifying intra-population gene-content diversity based on shotgun metagenomes, and we develop a tool, the **Gene Content Diversity (GCD) predictor** that implements the proposed pipeline. We hypothesized that the *de novo* metagenomic assembly of a natural population (consensus population genome) will encompass both core and flexible genes among the coexisting species members, which can be identified as genomic regions with differential coverage (sequencing depth) in read recruitment analysis because not all members of the population encode the variable genes by definition (Fig. 4.2). We first simulated metagenomic datasets with various levels of intra-population gene content diversity (multiple strains with differences in gene content), and evaluated the resulting

consensus population assemblies. Subsequently, by using a combination of skewed-log-normal distributions, we developed a method to model the variable sequencing depth across the length of the consensus population assemblies, and showed that it allows the reliable estimation of population gene content variability. We propose the **Core Content Index (CCI)**, a metric for the quantification of gene content diversity, and show that modeled sequencing depth variations can capture variations in CCI. Thus, given a metagenomic dataset and a *de novo* recovered population genome, the GCD-predictor can estimate gene content diversity in units of CCI. The method was evaluated in a separate sets of simulated metagenomes and resulted in typically small unbiased residuals. Finally, we applied the GCD-predictor to quantify intra-population gene content diversity in natural populations captured in metagenomic datasets from a variety of environments and contrasted the obtained values with CCI measured in isolate genome collections for model species.

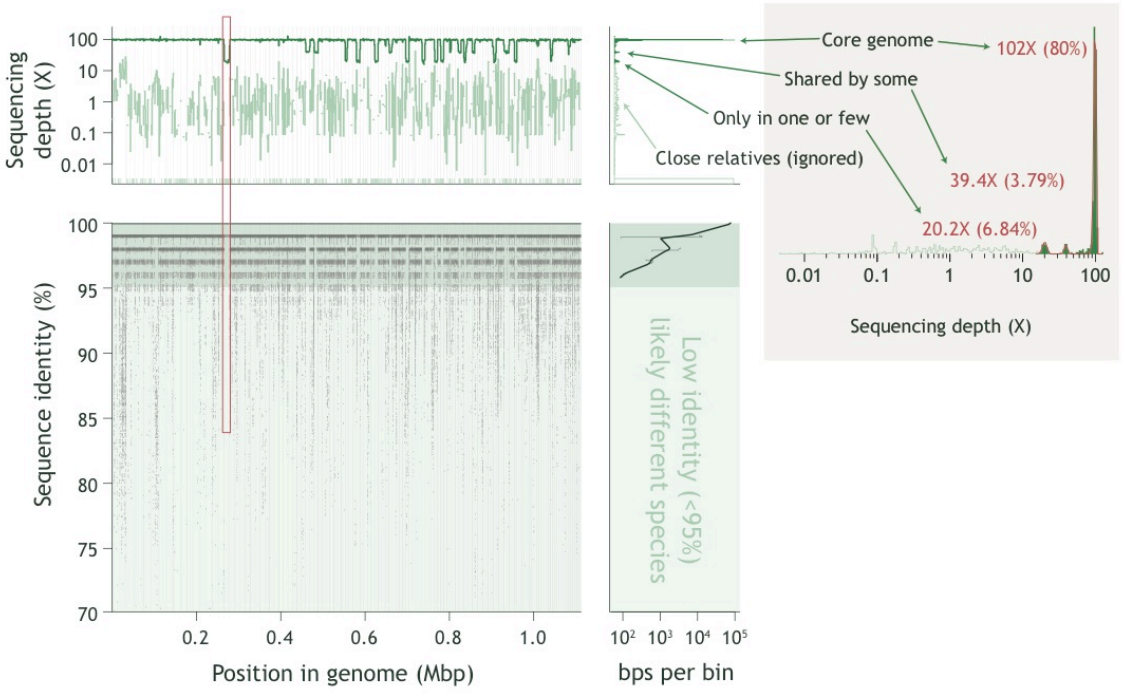


Figure 4.2: Example of metagenomic read recruitment analysis for the identification of gene content variation.

Bottom panel: Read recruitment against the *de novo* metagenomic assembly (consensus assembly of a population), where each dot represents a metagenomic read positioned along the genomic region where it was aligned (X-axis) and the nucleotide

sequence identity of the alignment (Y-axis). Reads >95% nucleotide identity typically represent discrete sequence clusters and are considered to belong to the same population. Some regions along the genome recruit fewer reads than the average, and are revealed in the histogram of sequencing depths (top panel), with an example region highlighted in the red rectangle. Quantification of such regions (representing flexible genes, shared only by some population members) relative to the total consensus genome, can provide a metric of relative flexible gene content. For this, the sequencing depth is projected into a histogram (right side of the top panel), and the relative area of the top peak (core genome) against the total areas represented by all the peaks (pangenome) is quantified. Thus the **relative core peak area** represents a metric of intra-population gene content diversity measured from the metagenome, using a *de novo* metagenomic assembly and read recruitment analysis. The recruitment plot presented here represents a simulation (see methods) in which simulated short reads from five *E. coli* closed genomes with gene content differences were spiked in a background metagenome, assembled *de novo* into a consensus *E. coli* population genome, which was subsequently used for the recruitment plot.

4.3 METHODS

4.3.1 Reference genome collections and estimation of pangenome

Isolate genomes from 10 model species with over 300 complete genomes available were collected from the NCBI Genome database (Table 4.1). For each species, we randomly subsampled ~300 genomes, in order to reduce the computational requirements for the subsequent analyses. Genes were predicted *de novo* using MetaGeneMark.hmm[25] to remove any systematic inconsistencies between multiple gene prediction methods. Reciprocal Best Matches (RBM) were identified between all pairs of genomes within each species collection using BLAST[26] and rbm.rb from the enveomics collection[27]. Orthology Groups (OG) were identified by combining all the RBM lists on each species collection using ogs.mcl.rb[27] and the distribution of OGs was used to determine the size of the pangenome (number of OGs), the core genome (number of OGs occurring at least once on each genome), and the average genome (average number of OGs per genome).

For each genome collection, Manhattan distances were calculated for each pair of genomes, as the fraction of different genes between two genomes over the average genome size of the species[28], and those distances were used in hierarchical clustering with the Ward method to construct the trees shown in Fig. 4.1.

Table 4.1: Statistics of the genome collections of the 10 named species used in this study.

Set	Species	Set size	Gene content statistics				Maximum flexible fraction[d]
			Ω[a]	C[b]	μ[c]	CCI	
AB	*Acinetobacter baumannii*	317	11,574	497	1,986	0.842	18.7%
BA	*Bacillus anthracis*	285	7,647	4,319	5,653	0.962	7.7%
CJ	*Campylobacter jejuni*	313	3,055	702	1,666	0.818	22%
EC	*Escherichia coli*[e]	364	16,282	505	4,648	0.903	29%
EF	*Enterococcus faecium*	318	8,275	488	2,584	0.808	36%
MT	*Mycobacterium tuberculosis*	328	5,379	1,767	3,966	0.948	7.2%
NM	*Neisseria meningitidis*	308	4,048	256	1,986	0.842	19%
SA	*Staphylococcus aureus*	313	5,177	1,122	2,749	0.858	25%
SP	*Streptococcus pneumoniae*	303	4,607	864	2,031	0.814	20%
VC	*Vibrio cholerae*	274	9,562	461	3,541	0.889	34%

a. Pangenome size (in OGs). b. Core genome size (in OGs). c. Average genome size (in OGs). d. Non-shared fraction of the genome between the most dissimilar pair of genomes in the collection in terms of gene content. e. Including *Shigella flexneri*.

4.3.2 Metrics of gene content diversity from reference genomes

In addition to the ratios Core/Pangenome, Average/Pangenome, and Core/Average, we propose a more robust metric of gene content variation derived from the OG distribution termed Core Content Index (Fig. 4.3):

$$CCI = 1 - \mathbb{E}_{i \in \Omega}\left[2 \times IC\left(f_g(i)\right)\right] = 1 + \frac{2 \times \sum_{i \in \Omega}\left[f_g(i) \times \log\left(f_g(i)\right)\right]}{|\Omega|}$$

Where *Ω* is the pagenome set (collection of all OGs) and $f_g(i)$ is the fraction of genomes containing the OG *i*. CCI is equal to 1 minus the expectation ($\mathbb{E}$) of two times the information content (*IC*) of the genome frequency of all OGs. All metrics were evaluated in (a) the 10 collections of 300 closed genomes per species and (b) 100 collections of 5 genomes per species in **metagenomic simulations** (see bellow), with various levels of gene content differences among the genomes used. The CCI index diversity metric better captures both the richness and evenness of OGs for a given collection, thus can more accurately represent gene content diversity in most cases (Fig. 4.3).

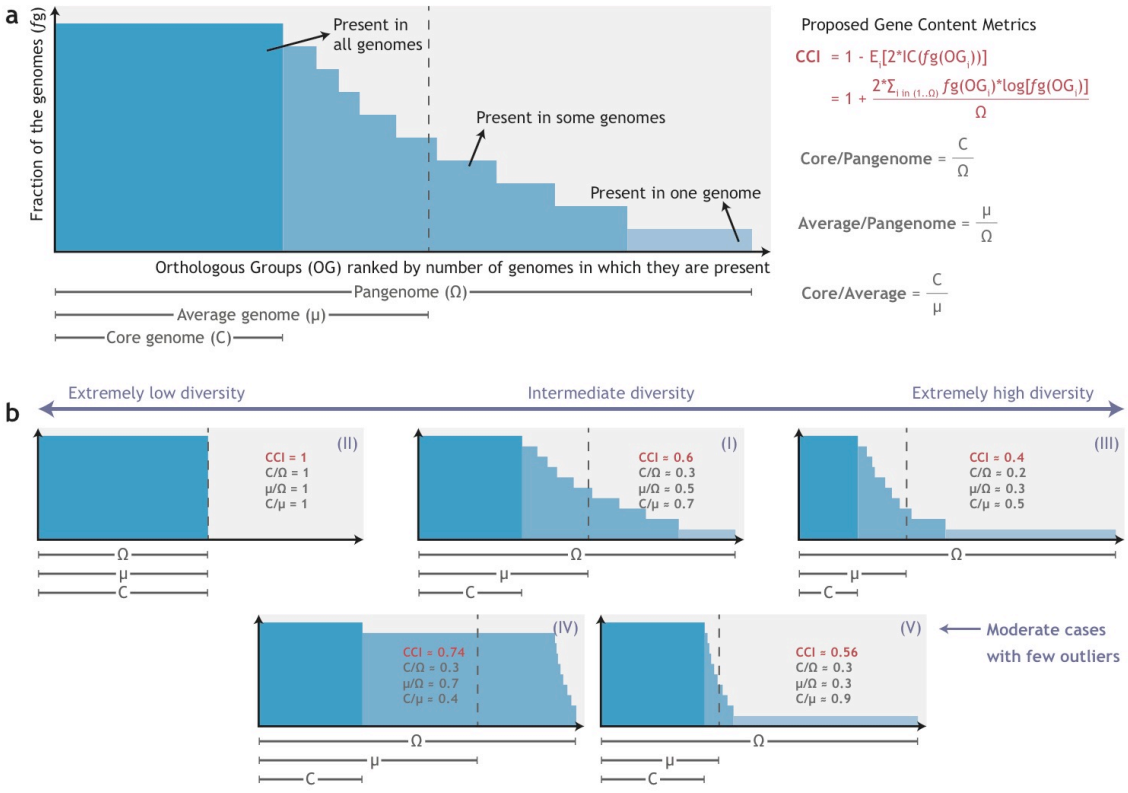


Figure 4.3: Gene content diversity metrics for a given set of genomes of the same species.

A. Conceptual illustrations of an OG distribution denoting the pangenome, average genome, and core genome in a collection of genomes from the same species. B. Variations of the OG distribution with different degrees of gene content diversity. Both the number of flexible genes as well as their representation in different genomes affect the observed gene content diversity in a given OG distribution. Case I represents an intermediate level of diversity, identical to A. Cases II and III represent extreme cases with no or high gene content diversity, respectively. Note that the level of diversity in cases I-III is captured by any of the proposed metrics. Cases IV and V represent genome collections with low-diversity distributions (between cases I and II), with a large deletion (IV) or a large insertion (V) in a subset of the genomes. Note that both changes should moderately increase the diversity at the same values, without reaching an extreme high diversity estimate (III). However, only CCI effectively captures this scenario without inflating the gene content diversity value.

4.3.3 Read recruitment analysis and sequence coverage modeling

For a given metagenomic read mapping onto a reference genome, sequencing depth was estimated for each gene (alternatively, it can be estimated for windows of fixed length), by counting the number of metagenomic reads aligned with >95% nucleotide identity against the reference genome using BLAT[29]. The top 5% highest sequencing depth is masked to prevent spurious results from highly conserved regions between different species, recent horizontal gene transfer, or other spurious read mapping events (for example identical duplicated regions tend to assemble as one, and show inflated sequencing depth values in recruitment analysis).

The resulting set of sequencing depth values is then modeled as a collection of skewed-log-normal distributions as follows. (i) The mode of the sequencing depth is estimated using the Parzen method[30]. (ii) The parameters of the log-normal-distribution are estimated from all sequencing depth values close to the mode using the quantiles method[31,32] increasing the range of values until the parameters stabilize. (iii) Modeled values are removed and remaining values are modeled using the same technique from step i. Iterative estimation and refinement of the modeled mixed distribution is performed until the model captures all but 10 or fewer sequencing depth values. Each skewed-log-normal distribution in the mix is termed a peak. The fraction of the values derived from the core peak (putatively corresponding to the core genome) is hereafter termed **Relative Core Peak Area**. The relative core peak area linearly correlates with the CCI allowing its estimation (Fig. 4.4), and this correlation is implemented in the GCD predictor. The parameters of the correlation were estimated from a set of 1,000 simulations (training sets), derived from the 10 genome collections of named species (Table 4.1) and evaluated in 92 simulations (testing sets), one for each prokaryotic species in NCBI Genome with at least 5 complete genomes available (Table 4.2). Read recruitment analysis was implemented in the R package enveomics.R as previously described[27], together with ancillary functions for peak identification, core peak selection, and estimation of CCI.

4.3.4 Simulation of metagenomic datasets

For each of the 10 genomes collections of the named species used we simulated 100 metagenomic datasets containing five genomes from the same species, with various levels of gene content differences (randomly chosen). The Grinder software[33] was used to simulate short Illumina sequencing reads at 20X coverage (parameters: -md uniform 0.1 -mr "95 5" -rd "100 uniform 5"), for each of the five reference genomes, and subsequently the produced reads were combined together with reads from a background metagenome (which didn't contain any of the species used in the simulations, and was generated in Ref. [34]), in order to resemble a high complexity microbial community. Each simulated metagenome was assembled with IDBA-UD[35], genes were predicted with MetaGenemark[25] and the resulting contigs were taxonomically annotated using the MyTaxa algorithm[36]. Contigs belonging to the consensus population assembly of the *in silico* added species in each simulation were identified based on their MyTaxa taxonomic affiliations, and were used in read recruitment analysis. The collection of 1,000 simulations (100 per species) was used as the training set for the quantification of gene content diversity, *i.e.*, correlation of core content index as estimated based on the OG distributions of the 5 genomes added in each simulation with the identified relative peak area from the recruitment analysis (Fig. 4.4). In order to test the robustness of the method we generated additional metagenomic simulations, with data that were not included in the training set (testing datasets; Table 4.2). For this, we identified species with at least 5 closed genomes available from the NCBI Genome archive excluding the 10 model species (Table 4.2), and followed the same procedure described above.

Table 4.2: Testing dataset for the GCD-predictor.

92 genomic collections of named species (with at least 5 closed genomes available) were used to test the accuracy of the GCD-predictor.

ID	Species Name	ID	Species Name	ID	Species Name
Acp	*Acetobacter pasteurianus*	Cog	*Co. glutamicum*	Mym	*My. mycoides*
Aeh	*Aeromonas hydrophila*	Cop	*Co. pseudotuberculosis*	Myo	*My. bovis*
Baa	*Bacillus amyloliquefaciens*	Cou	*Co. ulcerans*	Myp	*My. pneumoniae*
Bac	*Bacillus cereus*	Crs	*Cronobacter sakazakii*	Neg	*Neisseria gonorrhoeae*
Baf	*Bacteroides fragilis*	Ehc	*Ehrlichia chaffeensis*	Pam	*Pasteurella multocida*
Bam	*Bacillus megaterium*	Enc	*Enterobacter cloacae*	Pap	*Paenibacillus polymyxa*
Bao	*Bacillus coagulans*	Enf	*Enterococcus faecalis*	Pra	*Propionibacterium acnes*
Bas	*Bacillus subtilis*	Era	*Erwinia amylovora*	Psa	*Pseudomonas aeruginosa*
Bia	*Bifidobacterium animalis*	Flp	*Flavobaceter psychrophilum*	Psp	*P. putida*
Bib	*Bi. bifidum*	Frn	*Francisella noatunensis*	Psy	*P. syringae*
Bil	*Bi. longum*	Frp	*F. philomiragia*	Rhe	*Rhizobium etli*
Bir	*Bi. breve*	Frt	*F. tularensis*	Ria	*Riemerella anatipestifer*
Bob	*Borrelia burgdorferi*	Fun	*Fusobacterium nucleatum*	Rip	*Rickettsia prowazekii*
Bop	*Bordetella pertussis*	Hai	*Haemophilus influenzae*	Rir	*R. rickettsii*
Bra	*Brucella abortus*	Hep	*Helicobacter pylori*	Sae	*Salmonella enterica*
Brm	*Br. melitensis*	Laa	*Lactobacillus rhamnosus*	Shf	*Shigella flexneri*
Brs	*Br. suis*	Lac	*La. casei*	Sim	*Sinorhizobium meliloti*
Bum	*Burkholderia mallei*	Lah	*La. helveticus*	Ste	*Streptococcus equi*
Bup	*Bu. pseudomallei*	Lal	*La. plantarum*	Std	*St. dysgalactiae*
But	*Bu. thailandensis*	Lii	*Listeria ivanovii*	Stg	*St. agalactis*
Caf	*Campylobacter fetus*	Lim	*Li. monocytogenes*	Sts	*St. suis*
Caj	*Cam. jejuni*	Mah	*Mannheimia haemolytica*	Stt	*St. thermophilus*
CaP	*Ca.* Portiera aleyrodidarum	Mea	*Methanosarcina mazei*	Sui	*Sulfolobus islandicus*
Chm	*Chlamydia muridarum*	Meb	*Methanosarcina barkeri*	Sus	*Su. solfataricus*
Chp	*Ch. pneumoniae*	Mes	*Me. sedula*	Vip	*Vibrio parahaemolyticus*
Chs	*Ch. psittaci*	Mya	*Me. abscessus*	Xai	*Xanthomonas citri*
Cht	*Ch. trachomatis*	Myb	*Me. bovis*	Xyf	*Xylella fastidiosa*
Clb	*Clostridium botulinum*	Myc	*Mycoplasma capricolum*	Yee	*Yersinia enterocolitica*

Table 4.2 continued

Clm	*Clavibacter michiganensis*	Mye	*My. genitalium*	Yep	*Yersinia pestis*
Cob	*Coxiella burnetii*	Myg	*My. gallisepticum*	Zym	*Zymomonas mobilis*
Cod	*Corynebacterium diphtheriae*	Myh	*My. hyopneumoniae*		

4.3.5 Collection of natural population genomes from metagenomes

In order to quantify gene content diversity among natural bacterial populations, we identified a set of population genomes (Table C1) from available metagenomic datasets representing different environments (Table 4.3). Those population genomes had been *in silico* recovered using various binning methods (metagenomic bins). To reduce biases due to the different methods, the recovered population genomes were evaluated for their completeness and contamination using the CheckM software[37], and only bins with less than 5% estimated contamination and >80% completeness were maintained for further analysis. The sequencing coverage across each bin's genome length was estimated by read recruitment analysis as described above, based on the same metagenome from which the bin had been isolated. Finally, only bins with ≥10X average sequencing depth and ≥1 Mbp length were retained, as we noticed that the gene content estimations were not accurate for population genomes with lower abundance and sequencing breadth.

In order to test weather the gene content diversity in a given population correlates with the allelic diversity, we estimated Single Nucleotide Polymorphisms (SNPs) for each bin: Bowtie2[38] was used to map the reads against the reference bin, SNPs were identified using the SAMtools mpileup (with parameters: -C50 -g -f)[39] and BCFtools and VCFtools call[40], and quantified in both the entire genomes or after excluding likely flexible genes (detected by GCD-predictor) using the script VCF.SNPs.rb[27].

Table 4.3: Collections of population genomes isolated from various environments.

Collection	*# Bins*	*Isolation source*	*Reference*
Nitrate bioreactor[a,l]	2	Sediment Incubations amended with nitrate	41
Cyanate bioreactor	5	Continuous flow bioreactor with thiocyanite (SCN)	42
Anammox	1	Anammox enrichment RU1	43
Thermophilic reactor	11	Anaerobic biogas reactor (CSTR)	44
Multiple amendments[b]	7	Soil enrichments with various compounds	45
Hospital	1	Biofilm at hospital's shower head	46
Bankia symbionts[c]	2	*Bankia setacea* gill microbiome (1 individual)	47
Olavius symbionts	3	Microbiome of *Olavius algarvensis* (1 individual)	48
Tick endosymbiont	1	Gonads of *Rhipicephalus turanicus* (100 ticks)	49
Human gut	45	Stool samples Danish and Spanish human individuals	50
Freshwater	77	Lake Lanier time series samples	51
Marine (BS)	62	Baltic Sea brackish water time series samples	52
Marine (OMZ)	9	Oceanic Oxygen Minimum Zone waters	53
Marine (GOM)	22	Gulf of Mexico vertical profile waters	54
Estuary sediments[ll]	16	Sediments from White Oak River estuary	55
Soils	7	Alaska tundra soil samples	56
Groundwater sediments	100	Groundwater sediments	57
Marine sediments	2	Coal bed samples	58

Table 4.3 continued

For all datasets sequencing was performed with Illumina technology, except for (a) Ion torrent (b) 454 (c) Illumina and 454. Additionally, assemblies used for the recovery of metagenomic bins originated from single samples except from (I) where a co-assembly of four time series samples was used, and (II) where two soil samples from adjacent sites were co-assembled.

4.4 RESULTS AND DISCUSSION

4.4.1 Benchmarking the Gene Content Diversity (GCD) predictor

1,000 simulated datasets derived from the 10 model species (Table 4.1) were used to parameterize the correlation between relative core peak area and core content index (Fig. 4.4). The derived function was $CCI \approx 0.371 \cdot RCPA + 0.629$, where RCPA is the relative core peak area. Note that the expression has two parameters, but only one free parameter since the function is forced to the point (CCI=1, RCPA=1). This function was evaluated in the testing set (92 species with ≥5 genomes available; Table 4.2) resulting in an observed error of <0.1 CCI in 95% of the cases evaluated and a SMSE (Square root of the Mean Square Error) of 0.07 (Fig. 4.5). Moreover, the residual error didn't significantly correlate (p-values > 0.3) with any evaluated variable, including assembly length (note that all population genome assemblies used were at least 1Mbp long), gene-content statistics (pangenome, core genome, and average genome sizes), estimated CCI, or the estimated fit error from the skewed-log-normal mix distribution.

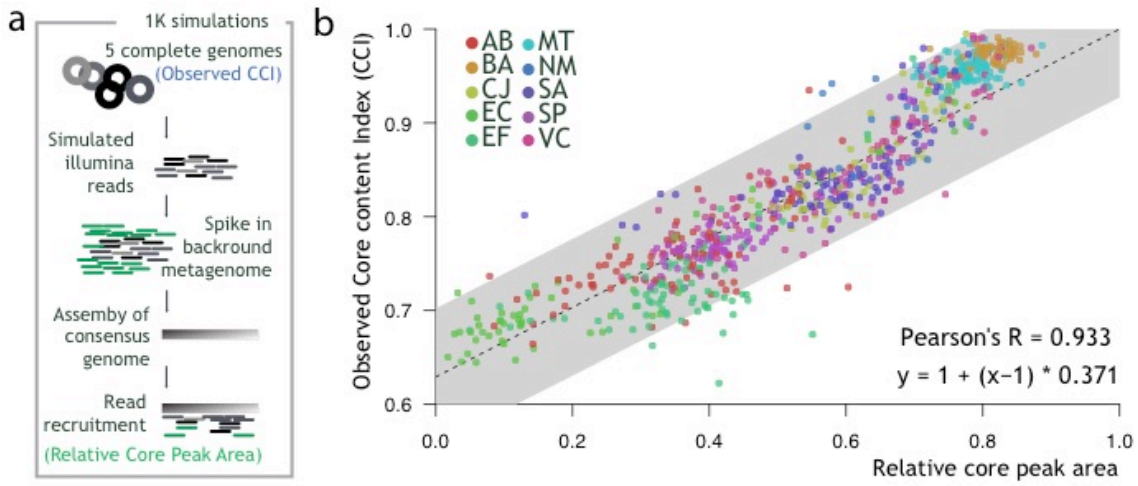


Figure 4.4: Method parameterization: Prediction of gene content variation in metagenomes.

10 genomic collections of named species (shown in Fig. 4.1) with at least 300 genome representatives were used to create *in silico* metagenomes, representing bacterial communities harboring between two and five genome representatives of the species of interest. **A.** For each species, 100 simulations were produced using 5 closed reference genomes per simulation. Data were assembled and contigs for the species of interest were identified and used in recruitment plot analysis. Relative core peak area was determined for each simulation from read recruitment analysis, and compared against the observed gene content diversity, estimated as CCI from the OG collection distribution from the original 5 genomes used in the simulation. **B.** The linear regression between Relative core peak area and CCI (dashed line) was determined by the least squares method with origin in (1,1) and free slope. The grey band indicates the area around the linear regression containing 95% of the data points.

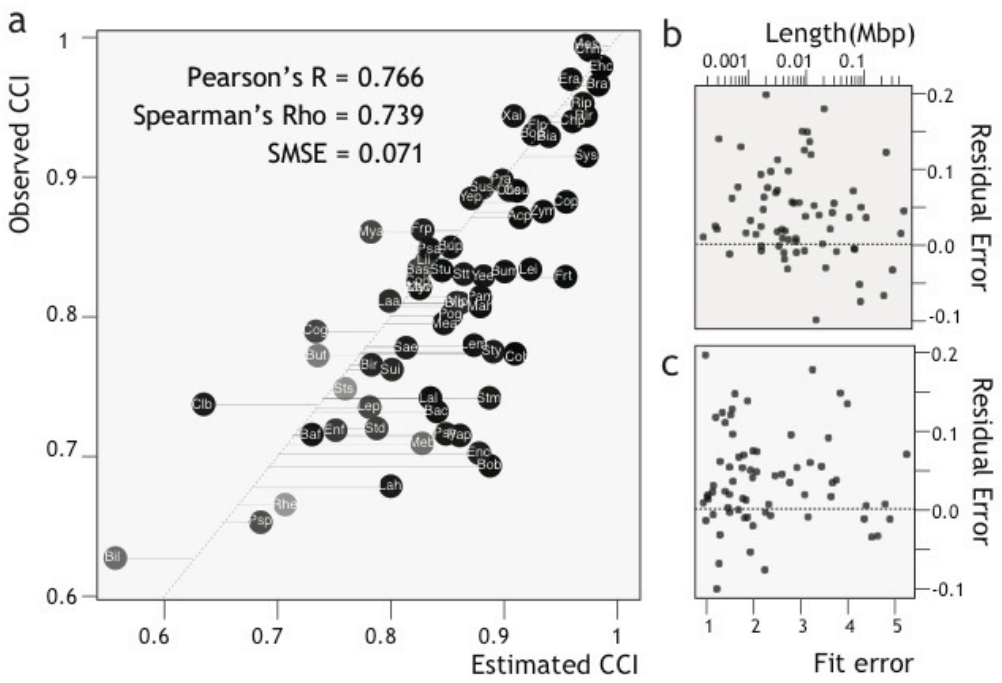


Figure 4.5: Method validation.

92 species with more than five closed genome representatives were used to evaluate the method. **A.** Linear correlation between estimated and observed CCI, including the Square root of the Mean Square Error (SMSE). Codes per datum correspond to those in Table 4.2. **B-C.** Residuals of the CCI estimation with spread by assembly length (A) and skewed-log-normal mix distribution fit error (B) showing no biases by either variable ($R < 0.02$). Residual error had an interquartile range of (-0.025, 0.025) and a 95% central interval of (-0.07, 0.1).

Thus, our method can reliably estimate gene content diversity from metagenomic data, provided a *de novo* assembled genome representative of the *in situ* population, and coverage of at least 10X for this genome in the metagenomic dataset. The implementation of this method has been incorporated in the **GCP predictor algorithm**.

4.4.2 Application of GCD-predictor in natural population data

We next examined the levels of gene content diversity harbored among natural bacterial populations, by applying the GCP predictor in multiple collections of population genomes (metagenomic bins, Table C1) that have been obtained from a variety of environments (Table 4.3). Overall, the results obtained were very similar across environments, with mean gene content diversity for the populations examined around 9.8 CCI (Fig 4.6). Such high CCI values are typical for highly clonal species, such as *B. subtillis* and *M. tuberculosis*, as shown in the reference CCI values estimated for all closed genome collections (rightmost column in Fig 4.6). Among the closed genome collections of *B. subtillis* and *M. tuberculosis*, the maximum gene content difference between two strains is ~7% of their total genes, and collectively the collections have CCI values ~0.95 (Table 4.3). For the natural populations, we cannot directly quantify the underlying OG distribution, however CCI values >0.95 represent minimal gene content diversity, lesser that the one observed among the named species. Additionally, those populations with high CCI values exhibited minimal variations at the sequence identity level, ie density of SNPs representing allelic diversity (see bellow). Thus, most of the natural populations examined here appear to be clonal in both dimensions of diversity, gene content and allelic, and observation expected for relative young populations that have merged (ie have lost their accumulated genetic diversity) from recent genetic sweeps[59]. Alternatively, the clonal nature of the populations could also be explained by the scenario in which those organisms are allochthonous species, ie have recently invaded the examined environment (metagenomic sample), or have recently emerged as abundant members from the rare biosphere, responding to environmental changes[17,60].

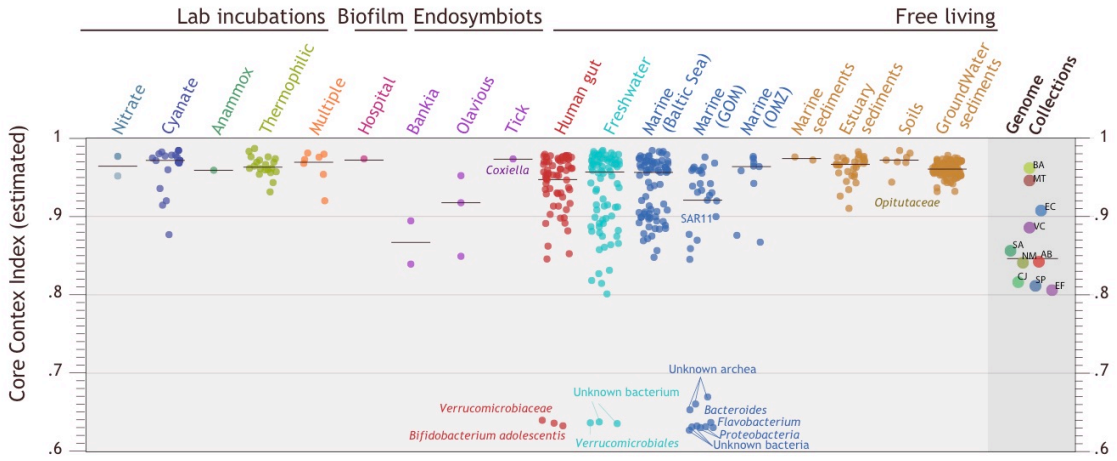


Figure 4.6: Gene-content diversity within natural bacterial populations.

Genomic bins were *in silico* isolated from multiple metagenomic datasets representing a variety of environments. For each genomic bin representing a natural bacterial population, the intra-population gene content variability was estimated from the recruitment plot histogram of the given bin and metagenomics dataset of origin. The results shown represent the estimated CCI from the recruitment plots for each bacterial population and populations were grouped by the source environment. The observed gene content variation of the whole available genome collection was calculated for 10 species, and is shown in terms of CCI in the rightmost column of the graph. Information on each of the collections of population genomes is provided in Table 4.3, and genomic statistics for each individual genomic bin in Table C1.

While most of the populations examined exhibited minimal intra-population diversity, significant levels of gene content diversity were detected for several atypical (outlier) populations, which originated mostly from environments with more data available (*i.e.*, Human gut, Freshwater, and Marine environments). Values of CCI varied from 1 down to 0.65, exceeding, in several cases, the gene content diversity captured by the collections of the reference species genomes. For some of those outlier populations with high gene content diversity we were able to identify several transposable elements and phages, which partially accounted for the difference in gene content. However, most of the variable genes captured by the GCD predictor showed similar functional distributions to the total genes of the population, *i.e.*, there was no significant functional enrichment detected (Fig. 4.7). This observation is consistent with the neutral model for maintaining most -but likely not all- of this gene content diversity within a sympatric population, as no apparent functional adaptations were identified[17]. However, it is likely that the flexible genes of those populations might include adaptive traits that our method and data were not able to recognize, and indeed a large fraction of those genes were unclassified or annotated as proteins of unknown function (Fig. 4.7). Furthermore, the taxonomic affiliations of the evaluated bins revealed a taxonomic bias in the gene content diversity of the populations: Population genomes belonging to the *Spirochaetes*, *Bacteroidetes*, or *Verrucomicrobia* phyla were enriched in the subset of genomes with low CCI values (high gene content diversity), and those results were consistent across diverse environments, such as human gut microbiome and aquatic ecosystems (Fig. 4.6). The consistently low CCI values for populations of the above phyla across habitats might reflect distinct modes of genomic evolution and diversity maintenance in the latter organisms.

We further evaluated the scenario that the observed gene content diversity might correlate with the accumulated allelic diversity expressed as the density of SNPs found in the population's genome. Interestingly, we didn't observe a significant correlation between estimated CCI and total genomic SNP density in the entire genomic assemblies (p-value 0.28; Fig. 4.8) or excluding putative flexible genes detected by GCD-Predictor (p-value 0.47). This is a surprising result given that both gene content diversity and SNP density are thought to be cumulative variables increasing with the population age as parallel dimensions of micro-diversity[61], *i.e.*, the population diversity accumulated since

its introduction or since the last strong genetic bottleneck. These results are consistent with the majority of the variable genes being neutral, as also revealed by the functional annotation analysis mentioned above, at least for the average (prevailing) conditions within the corresponding environment that were preferentially sampled here by the available metagenomes. It is reasonable to expect that some of these variable genes are functionally important when conditions fluctuate within the environment; however, the great majority of the metagenomes analyzed here were obtained under stable, not-fluctuating conditions and so, were not informative with respect to fluctuating conditions. Further, it is important to note that if an ecologically or functionally distinct sub-population has emerged only recently from within the population so that has not differentiated enough in terms of sequence diversity (to be detected in the recruitment plot as uneven coverage) or gene-content (to be detected by the CCI analysis), it would not be captured by our approach. Collectively, it appears that the natural populations analyzed here, even the ones with extensive intra-population gene-content diversity, meet several of the key attributes expected for "species" such as they are genetically discrete from co-occurring populations (e.g., genetic discontinuity areas in recruitment plots around 90-95%), have minimal or mostly natural intra-population diversity (CCI and functional analysis above), and members of the population show similar *in-situ* abundances under the same environmental conditions (i.e., same sample), reflected by even coverage in the recruitment plots. These findings further corroborated our previous results based on a much more limited dataset and assessment of intra-population gene content diversity that the natural sequence-discrete populations probably represent the most important units of microbial communities[60].

Several genetic and ecological mechanisms could account for the results observed. On one hand, gene content is largely shaped by gene acquisition through horizontal gene transfer, gene duplication and gene loss though genome streamlining and genetic drift[15,17,62], and both gene loss and acquisition are mediated by environmental selection pressure, which might fluctuate as the growth conditions fluctuate. On the other hand, SNP density increases as a neutral process, but is also affected by phenomena such as homologous recombination[63], heterogeneous mutation rates[64], genetic sweeps[65], and incomplete lineage sorting[66]. The lack of statistically significant correlation between SNP density and estimated CCI underscores the

importance of estimating both dimensions of micro-diversity in the characterization of natural populations. Therefore, GCD-Predictor represents an important complement to SNP density or core-genome phylogeny for the estimation of genetic population diversity and for future studies that will attempt to quantify the relative contribution of the different genetic and ecological mechanisms based on natural populations.

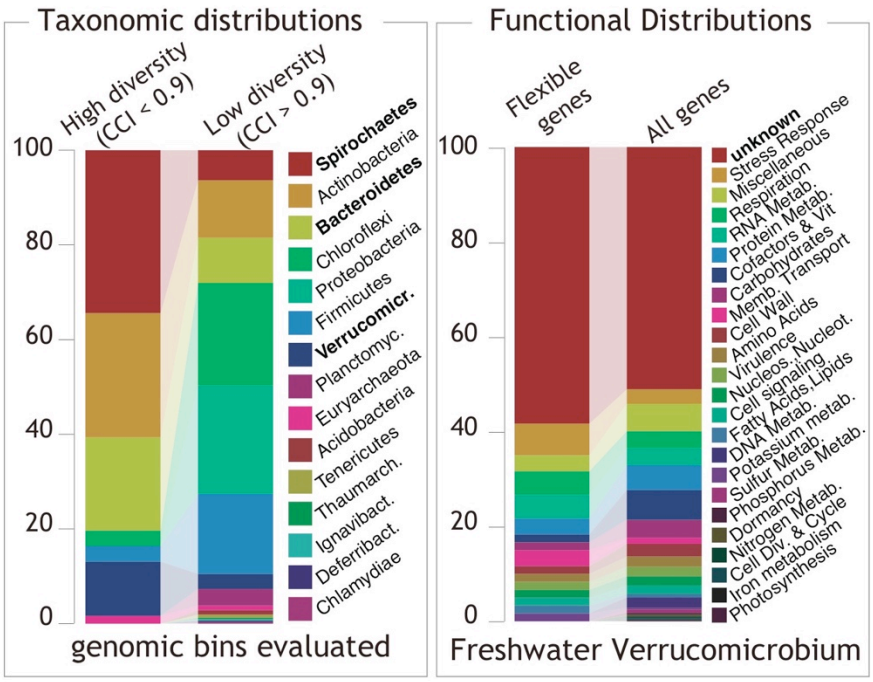


Figure 4.7: Functional and taxonomic distributions in populations with high gene content diversity.

Taxonomic distribution (left panel) of all population genomes that were evaluated with the GCD predictor (left column). Populations identified as outliers with high gene content diversity are predominantly belonging to the *Spirochaetes*, *Bacteroidetes*, and *Verrucomicrobia* phyla (right column). A typical example of a functional distribution is shown of one of the outlier populations, a freshwater *Verrucomicrobium* with CCI value of 6.9. There is no strong functional enrichment among the flexible genes detected, but about 50% of the total genes of this population, as it is the case for most other populations with high gene content diversity, are of unknown function or unclassified.

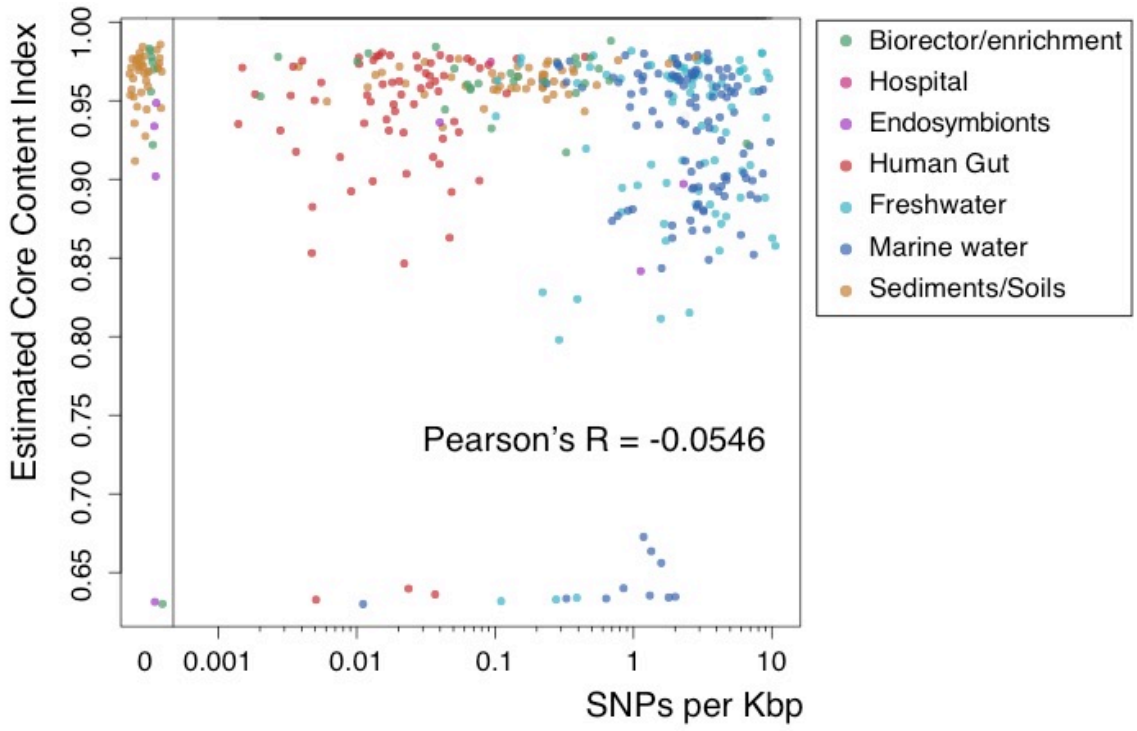


Figure 4.8: Correlation of gene content with allelic diversity among natural populations.

For each population genome shown in Fig 4.7, gene content diversity was estimated in terms of CCI using the GCD-Predictor, and allelic diversity was estimated as the frequency of SNPs among members of (i.e., sequences assigned to) the population. No significant correlation was detected between estimated CCI and the SNP density in the entire genomes (R: -0.055, p-value: 0.28) or in genes excluding putative flexible genes (R: -0.036, p-value: 0.47, data not shown). Note that allelic diversity values are presented in logarithmic scale. Population genomes with no detected SNPs are shown in the leftmost box with random noise added to visualize overlapping values.

4.5 CONCLUSIONS

The extent and nature of microdiversity within bacterial populations in terms of allelic and gene content variations can unravel the underlying evolutionary mechanisms of how microbial populations evolve and adapt to their environments. Additionally, this information might be crucial for determining the value of biodiversity for microbial community functioning and resilience. Metagenomic datasets across environments offer the opportunity to assess both dimensions of microdiversity within populations *in situ* and quantify their importance in evolutionary and ecological processes. Here we present GCD-Predictor, a tool that can quantify gene content diversity of natural populations from metagenomes. Our method relies on the availability of a representative population genome (metagenomic bin) with ≥10X sequencing depth to estimate the gene content variations at the sub-species level. Application of this method to available natural bacterial populations revealed that most bacterial populations have low gene content differences within the same environment, an observation applicable to a variety of habitats. Nevertheless, our method also detected a number of outliers to this pattern, typically from the aquatic bacteria from the *Spirochaetes*, *Bacteroidetes*, and *Verrucomicrobia* phyla, which appear to accumulate extensive amount of gene content intra-population diversity. Most -but likely not all- of the latter diversity appears to be neutral, at least for the prevailing conditions within the corresponding habitat, since it was not significantly enriched in specific gene functional categories (e.g., Fig. 4.7) and did not vary in genotypes of the population that were more divergent at the sequence identity (SNPs) level relative to more closely related genotypes (e.g., Fig. 4.8). Extending this analysis to additional populations across habitats and evaluating the nature of the identified flexible genes might provide further quantitative insights into the interplay of ecologic and evolutionary processes *in situ* and the ecological mechanisms underlying these patterns of diversity.

4.6 REFERENCES

1. Welch, R. A. *et al.* Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proc. Natl. Acad. Sci.* **99,** 17020–17024 (2002).
2. Parkhill, J. *et al.* The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences. *Nature* **403,** 665–668 (2000).
3. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial 'pan-genome'. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 13950–13955 (2005).
4. Wu, K.-M. *et al.* Genome sequencing and comparative analysis of Klebsiella pneumoniae NTUH-K2044, a strain causing liver abscess and meningitis. *J. Bacteriol.* **191,** 4492–4501 (2009).
5. Chun, J. *et al.* Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic Vibrio cholerae. *Proc. Natl. Acad. Sci.* **106,** 15442–15447 (2009).
6. Rouli, L., Merhej, V., Fournier, P.-E. & Raoult, D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect.* **7,** 72–85 (2015).
7. Chan, A. P. *et al.* A novel method of consensus pan-chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of Acinetobacter baumannii. *Genome Biol.* **16,** 143 (2015).
8. Luo, C. *et al.* Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. *Proc. Natl. Acad. Sci.* **108,** 7200–7205 (2011).
9. Hugh-Jones, M. & Blackburn, J. The ecology of Bacillus anthracis. *Mol. Aspects Med.* **30,** 356–367 (2009).
10. Namouchi, A., Didelot, X., Schöck, U., Gicquel, B. & Rocha, E. P. C. After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection. *Genome Res.* **22,** 721–734 (2012).
11. Smith, N. H., Gordon, S. V., de la Rua-Domenech, R., Clifton-Hadley, R. S. & Hewinson, R. G. Bottlenecks and broomsticks: the molecular evolution of Mycobacterium bovis. *Nat. Rev. Microbiol.* **4,** 670–681 (2006).
12. Bolotin, E. & Hershberg, R. Gene Loss Dominates As a Source of Genetic Variation within Clonal Pathogenic Bacterial Species. *Genome Biol. Evol.* **7,** 2173–2187 (2015).
13. McConnell, M. J., Actis, L. & Pachón, J. Acinetobacter baumannii: human infections, factors contributing to pathogenesis and animal models. *FEMS Microbiol. Rev.* **37,** 130–155 (2013).
14. Kashtan, N. *et al.* Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus. *Science* **344,** 416–420 (2014).
15. Wilmes, P., Simmons, S. L., Denef, V. J. & Banfield, J. F. The dynamic genetic repertoire of microbial communities. *FEMS Microbiol. Rev.* **33,** 109–132 (2009).
16. Biller, S. J., Berube, P. M., Lindell, D. & Chisholm, S. W. Prochlorococcus: the structure and function of collective diversity. *Nat. Rev. Microbiol.* **13,** 13–27 (2015).

17. Cordero, O. X. & Polz, M. F. Explaining microbial genomic diversity in light of evolutionary ecology. *Nat. Rev. Microbiol.* **12,** 263–273 (2014).
18. Fraser, C., Alm, E. J., Polz, M. F., Spratt, B. G. & Hanage, W. P. The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity. *Science* **323,** 741–746 (2009).
19. Konstantinidis, K. T. & DeLong, E. F. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J.* **2,** 1052–1065 (2008).
20. Doolittle, W. F. & Papke, R. T. Genomics and the bacterial species problem. *Genome Biol.* **7,** 116 (2006).
21. Thompson, J. R. *et al.* Genotypic Diversity Within a Natural Coastal Bacterioplankton Population. *Science* **307,** 1311–1313 (2005).
22. Coleman, M. L. *et al.* Genomic Islands and the Ecology and Evolution of Prochlorococcus. *Science* **311,** 1768–1770 (2006).
23. Konstantinidis, K. T., Ramette, A. & Tiedje, J. M. The bacterial species definition in the genomic era. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361,** 1929–1940 (2006).
24. Wilmes, P. *et al.* Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. *ISME J.* **2,** 853–864 (2008).
25. Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* **38,** e132 (2010).
26. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
27. Rodriguez-R, L. M. & Konstantinidis, K. T. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. (2016). doi:10.7287/peerj.preprints.1900v1
28. Murtagh, F. & Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J. Classif.* **31,** 274–295 (2014).
29. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
30. Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **33,** 1065–1076 (1962).
31. Klugman, S. A., Panjer, H. H. & Willmot, G. E. *Loss models: from data to decisions*. (Wiley, 2012).
32. Delignette-Muller, M. L. & Dutang, C. fitdistrplus: An R Package for Fitting Distributions. *J. Stat. Softw.* **64,** 1–34 (2015).
33. Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P. & Tyson, G. W. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.* gks251 (2012). doi:10.1093/nar/gks251
34. Rodriguez-R, L. M. & Konstantinidis, K. T. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* **30,** 629–635 (2014).
35. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinforma. Oxf. Engl.* **28,** 1420–1428 (2012).
36. Luo, C., Rodriguez-R, L. M. & Konstantinidis, K. T. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Res.* gku169 (2014). doi:10.1093/nar/gku169

37. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25,** 1043–1055 (2015).
38. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
39. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
40. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).
41. Kraft, B. *et al.* The environmental controls that govern the end product of bacterial nitrate respiration. *Science* **345,** 676–679 (2014).
42. Kantor, R. S. *et al.* Bioreactor microbial ecosystems for thiocyanate and cyanide degradation unravelled with genome-resolved metagenomics. *Environ. Microbiol.* **17,** 4929–4941 (2015).
43. Speth, D. R. *et al.* Comparative genomics of two independently enriched 'Candidatus Kuenenia stuttgartiensis' anammox bacteria. *Evol. Genomic Microbiol.* **3,** 307 (2012).
44. Campanaro, S. *et al.* Metagenomic analysis and functional characterization of the biogas microbiome using high throughput shotgun sequencing and a novel binning strategy. *Biotechnol. Biofuels* **9,** (2016).
45. Delmont, T. O. *et al.* Reconstructing rare soil microbial genomes using in situ enrichments and metagenomics. *Syst. Microbiol.* **6,** 358 (2015).
46. Soto-Giron, M. J. *et al.* Characterization of biofilms developing on hospital shower hoses and implications for nosocomial infections. *Appl. Environ. Microbiol.* AEM.03529-15 (2016). doi:10.1128/AEM.03529-15
47. O'Connor, R. M. *et al.* Gill bacteria enable a novel digestive strategy in a wood-feeding mollusk. *Proc. Natl. Acad. Sci.* **111,** E5096–E5104 (2014).
48. Seah, B. K. B. & Gruber-Vodicka, H. R. gbtools: Interactive Visualization of Metagenome Bins in R. *Microb. Physiol. Metab.* 1451 (2015). doi:10.3389/fmicb.2015.01451
49. Gottlieb, Y., Lalzar, I. & Klasson, L. Distinctive Genome Reduction Rates Revealed by Genomic Analyses of Two Coxiella-Like Endosymbionts in Ticks. *Genome Biol. Evol.* **7,** 1779–1796 (2015).
50. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32,** 822–828 (2014).
51. Tsementzi, D., Poretsky, R., Rodriguez-R, L. M., Luo, C. & Konstantinidis, K. T. Evaluation of metatranscriptomic protocols and application to the study of freshwater microbial communities. *Environ. Microbiol. Rep.* **6,** 640–655 (2014).
52. Hugerth, L. W. *et al.* Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biol.* **16,** 279 (2015).
53. Ganesh, S., Parris, D. J., DeLong, E. F. & Stewart, F. J. Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. *ISME J.* **8,** 187–211 (2014).
54. Tsementzi, D., Meziti, A. & Konstantinidis, K. Metagenomic insghts into the adaptations of microbial life in the deep see ocean. *Prep.*
55. Baker, B. J., Lazar, C. S., Teske, A. P. & Dick, G. J. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. *Microbiome* **3,** 14 (2015).

56. Johnston, E. R. *et al.* Metagenomics Reveals Pervasive Bacterial Populations and Reduced Community Diversity across the Alaska Tundra Ecosystem. *Front. Microbiol.* **7,** 579 (2016).
57. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523,** 208–211 (2015).
58. Evans, P. N. *et al.* Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science* **350,** 434–438 (2015).
59. Bendall, M. L. *et al.* Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J.* **10,** 1589–1601 (2016).
60. Caro-Quintero, A. & Konstantinidis, K. T. Bacterial species may exist, metagenomics reveal. *Environ. Microbiol.* **14,** 347–355 (2012).
61. Schloter, M., Lebuhn, M., Heulin, T. & Hartmann, A. Ecology and evolution of bacterial microdiversity. *FEMS Microbiol. Rev.* **24,** 647–660 (2000).
62. Lynch, M. *et al.* Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* **17,** 704–714 (2016).
63. Go, M. F., Kapur, V., Graham, D. Y. & Musser, J. M. Population genetic analysis of Helicobacter pylori by multilocus enzyme electrophoresis: extensive allelic diversity and recombinational population structure. *J. Bacteriol.* **178,** 3934–3938 (1996).
64. Wielgoss, S. *et al.* Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc. Natl. Acad. Sci.* **110,** 222–227 (2013).
65. Cohan, F. M. & Perry, E. B. A Systematics for Discovering the Fundamental Units of Bacterial Diversity. *Curr. Biol.* **17,** R373–R386 (2007).
66. Degnan, J. H. & Rosenberg, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* **24,** 332–340 (2009).

CHAPTER 5

INSIGHTS INTO THE MUTUALISTIC ASSOCIATIONS OF *Dehalococcoides mccartyi* WITHIN DECHLORINATING MICROBIAL CONSORTIA

5.1 ABSTRACT

Dehalococcoides mccartyi (*Dhc*)'s energy metabolism is restricted to reductive dechlorination, and, based on its metabolic properties, this organism plays a keystone role in bioremediation of chlorinated solvent pollutants such as Trichloroethene (TCE) and Vinyl Chloride (VC). The efficiency of reductive dechlorination depends on (a) interactions with co-occurring microbial community members, and (b) the presence of *Dhc* strains with the appropriate reductive dehalogenase gene repertoire required for each specific contaminant. Understanding the nature of those interactions will inform the design and monitoring of successful bioremediation or bio-augmentation efforts. Here, we recovered 14 nearly complete genomes from coexisting community members in *Dhc*-containing dechlorinating mesocosms. Several of these genomes accounted for up to 80% of the total bacteria in enriched TCE-dechlorinating communities. Analysis of time series metranscriptomes of the enriched community during the course of dechlorination identified distinct strains of *Dhc* as major dechlorintors in this system, along with other TCE dechlorinators such as *Geobacter lovleyi, Sulfurospirillum* and *Desulfitobacterium* species. All identified community members exhibited overlapping functional roles, with varying contributions throughout the course of dechlorination. Among them, the identified dechlorinator species were the most predominant members at the offset of the dechlorination activity, and their gene expression dynamics indicated a role in the production of essential cofactors required by *Dhc*. Taken all together, our results

identified overlapping and dynamic functional interactions among community members that likely underlie the robustness of dechlorinating mixed cultures.

5.2 INTRODUCTION

Chlorinated pollutants comprise more than 50% of the top 100 pollutants in the 2015 Substance Priority list based on the Agency for Toxic Substances & Disease Registry (http://www.atsdr.cdc.gov/SPL/index.html). Chlorinated compounds are naturally produced from a variety of processes like combustion, volcanic activity, chemical and photochemical processes[1]. However, high local concentrations are typically a result of anthropogenic activities. Among those compounds, chlorinated ethenes and particularly PCE (perchloroethene), TCE (trichloroethene) and VC (vinyl chloride) are **prevalent groundwater contaminants**, present in more than 50% of the US EPA National Priority List sites[2]. **Reductive dechlorination** (Fig. 5.1) is the primary mechanism for the *in situ* biodegradation of chlorinated ethenes **under anoxic conditions**[3–5], in which the chlorinated ethenes are used as an electron acceptor for energy gain by specific microbial species termed "dechlorinators"[5,6]. While several different organisms from divergent phyla are able to dechlorinate PCE and TCE[6,7], complete dechlorination has only been observed in strains of *Dehalococcoides* (*Dhc*), which are capable of reductive dechlorination of 1,2 cis dichloroethene (cDCE) and vinyl chloride (VC) to the non toxic ethene[8–11].

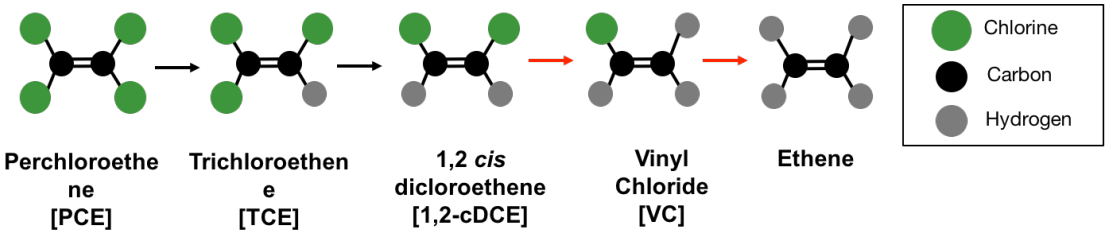


Figure 5.1: Reductive dechlorination of PCE.

PCE is sequentially dechlorinated to TCE, cDCE (cis-dechloroethene), VC and finally ethene. *Dhc* strains are the only known dechlorinators up to date that can perform the last two steps (red arrows).

All *Dehalococcoides* strains isolated to date belong to the single species *Dehalococcoides mccartyi* [12], and they are all highly specialized dechlorinators. Their genomes are among the smallest ever reported for free living organisms (~1.5Mbp) and reflect their **highly restricted lifestyle** [13]**:** they can only grow **by reductive dechlorination**, using hydrogen as electron donor and acetate as a carbon source, and no other growth-supporting redox couples have been identified [11,14–16]. Individual *Dhc* genomes contain only the essential metabolic pathways and up to **36 genes that encode for reductive dehalogenases** (**RDases**), the key enzymes in reductive dechlorination [13,17,18]. The repertoire of RDases in each genome dictates the specificity of *Dhc* strains for various chlorinated compounds, and not all the strains have the ability of complete dechlorination [19]. Apart from the various RDases, all *Dhc* strains have nearly identical genomes, with several essential biosynthetic pathways lacking or being incomplete [11]. For example, all known strains are corrinoid auxotrophs, even though a vitamin B_{12} derivative is essential for catalyzing reductive dechlorination reactions[20,21].

As a consequence, all isolates are fastidious growers. Their maintenance as axenic cultures is extremely difficult and thus most *Dhc* are typically maintained in mixed cultures [22]. Additionally, within mixed cultures they exhibit higher growth rates and stability and more efficient dechlorination [23–27]. The non-dechlorinating members of the mixed communities are typically fermentative, acetogenic bacteria and methanogens [28–31]. Although some members compete with *Dhc* for electron donors, some others benefit *Dhc* by providing hydrogen and carbon source (acetate), through net transformation of soluble organic electron donors in mixed cultures [24,32]. However, when pure *Dhc* cultures are provided with sufficient hydrogen, acetate and essential cofactors (*e.g.*, vitamin B_{12}), they don't grow as robustly compared to the mixed cultures [10,27]. Thus, community members appear to offer additional, yet-to-be-discovered co-factors or services (e.g., scavenging oxygen radicals) that affect the *Dhc* growth and dechlorination stability and activity.

It is expected, and reported in some cases, that microbial community members sustain *Dhc* growth by providing additional essential cofactors and metabolites which *Dhc* cannot synthesize [33,34]. Another possibility is that specific community members are providing protection from oxidative stress, which is essential for the strictly anaerobic lifestyle of *Dhc* [31]. Consequently, during isolation efforts of *Dhc* strains, dechlorination rates decrease, lag times for the onset of reductive dechlorination and associated growth

increase, and *Dhc* cell yields per mol of chloride released decrease with increasing culture purity[27,35]. In a few cases, co-culture experiments of *Dhc* with associate members isolated from the same sites have shown partner organisms are providing vitamin B_{12}, essential for dechlorination activity [25,33], which *Dhc* cannot synthesize *de novo* [20,21]. However, in most cases the role of the community members remains speculative [31,32,34,36,37]. Understanding the nature of microbe-microbe interactions within well-defined and robust dechlorinating microbial consortia will be valuable for monitoring, predicting and enhancing the bioremediation of chlorinated organic compounds in the environment.

In this work we aimed to characterize key community members within an established TCE-dechlorinating community, in order to identify the microbial interactions that affect *Dhc* and subsequently the course of dechlorination. A sediment free enriched TCE-dechlorinating community was obtained by serial transfers from actively dechlorinating mesocosms that were established using contaminated sediments described in a previous study [38]. Here, we sequenced the metagenome of the enriched community, along with metagenomes from the original mesocosms and, by using high-coverage data, we were able to recover nearly complete genomes of the predominant community members. Finally, we monitored the highly adapted TCE-dechlorinating enrichment with time series metatranscriptomes through the course of complete dechlorination. Genomic analysis of the recovered genomes and their relative transcriptional contributions in the mixed community enabled the identification of potential roles and interactions that sustained the dechlorinating activity in the mixed community.

5.3 METHODS

5.3.1 Establishment of the TCE enrichment culture

A sediment-free TCE-dechlorinating enrichment culture has been maintained in Löffler's lab for the past 3 years. It was originally established with sediments from a tributary of the Tennessee River in Knoxville, TN (Site 2, Third Creek site, Fig 5.2), with a well reported history of localized contamination with chlorinated ethenes [38]. In brief, the original mesocosms were maintained in batch cultures in anoxic bicarbonate-buffered,

defined mineral salts medium [39] amended with 0.4 mM TCE (99%, from Fisher Scientific, Pittsburgh, PA) as electron acceptor, and 5 mM lactate as electron donor and incubated at 30°C in the dark. The mesocosms were allowed to complete the reductive dechlorination of TCE, cis-DCE and VC, and were subsequently transferred to fresh medium by 3% v/v. After 7 transfers within a two-year period, sediment-free enriched cultures were established that were able to completely dechlorinate TCE to ethene. This enrichment culture is hereafter referred as E3 enrichment. For this study, ten replicated cultures from the E3 enrichment were prepared with 2% v/v inoculum in fresh minimal salts medium, 0.4 mM TCE and 5 mM of lactate, and were routinely monitored for total bacterial and *Dhc* 16S rRNA gene counts, concentrations of chlorinated ethenes, and lactate fermentation products, until complete dechlorination was achieved after 77 days. Replicated cultures were sacrificed for total RNA extractions at 4 time points (see below) throughout the experiment.

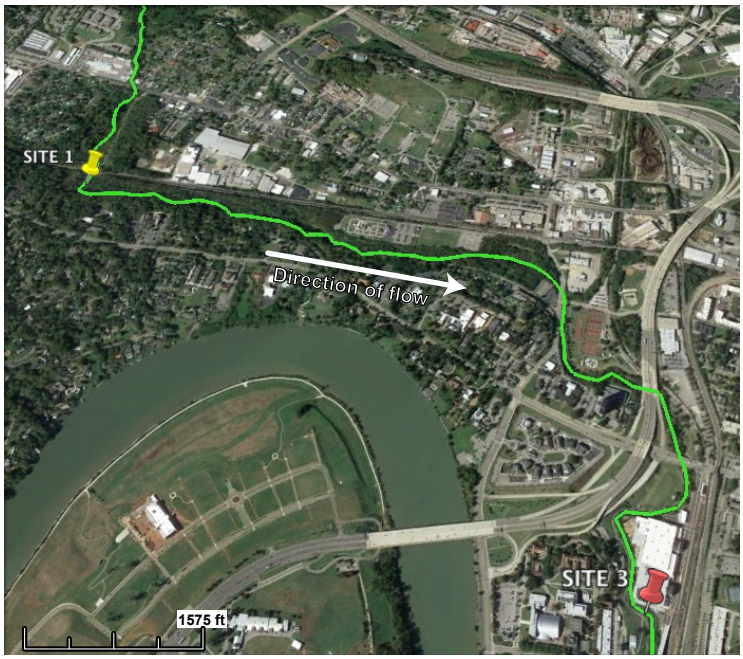


Figure 5.2: Location of the sampling sites within Third Creek.

Location of the contaminated creek where the streambed is infiltrated with a mixture of various chlorinated solvents (localized contamination with chloroethenes, Site 3). Site 1 represents an upstream sampling site with no history of contamination with chloroethenes. Mesocosms with sediments from the contaminated site with various electron acceptors (PCE, TCE and cDCE) and lactate that perform complete dechlorination have been previously shown to be dominated by *Dhc* strains [40].

5.3.2 Analytical techniques

Chlorinated ethenes and ethene were monitored using an Agilent 7890 gas chromatograph (GC) equipped with a flame ionization detector and a DB-624 capillary column (60 m by 0.32 mm with a film thickness of 1.18 μm) as previously described [41]. The method provided linear detector responses for quantification of each analyte to 0.7 mM aqueous phase concentrations with a detection limit of about 2.5 μM. Lactate and its fermentation products acetate and propionate were analyzed using an Agilent 1200 series HPLC system equipped with and Aminex HPX-97H column (Bio-Rad, Hercules, CA, USA). Samples were acidified with 1 M H_2SO_4 in a ratio of 19:1 (v/v) and separated with 4 mM aqueous H_2SO_4 as the mobile phase at a flow rate of 0.6 ml/min and quantified using a UV detector set to 210 nm.

5.3.3 DNA and RNA extractions and sequencing

DNA extractions were performed for six samples in total: (a) 2 sediment samples from the Third Creek site, collected 2 years apart, (b) 3 samples from active mesocosms that have been established from the contaminated sediments with either PCE, TCE or cDCE, and (c) from the sediment free enrichment culture E3, which originated from the TCE-dechlorinating mesocosm, after 7 sequential transfers (Table 5.1). The extractions were performed using 5 grams of soil from the original sediments, or 100 ml of suspended cells from the mesosocosms or the enrichment, when all the chlorinated substrates had been fully degraded. For the E3 enrichment, 1 ml samples were periodically sampled to be used for qPCR assays. A replicate of the culture was finally sacrificed for a DNA extraction at the end of the dechlorination (day 77), in order to obtain sufficient material for metagenomic library preparation. For all DNA samples extractions were performed with the MO Bio Soil DNA Isolation kit (MO BIO, Carlsbad, CA, USA). For the E3 enrichment total bacterial, and *Dhc* 16S rRNA gene copies were quantified by quantitative PCR (qPCR) using Bac1055F/Bac1392R/Bac1115Probe and Dhc1200F/1271R/1240Probe primer-probe sets following the established qPCR protocol [42].

For the RNA extractions, approximately 100 ml triplicate TCE-dechlorinating cultures were sacrificed at four different time points (5^{th}, 7^{th}, 28^{th} and 77^{th} day). The same triplicate cultures were also sampled for 16S rRNA qPCR, organic acid and chlorinated ethene measurements. Cultures were transferred into 50-mL sterile plastic tubes inside

an anoxic glove box filled with H_2/N_2 (3/97%, v/v) and centrifuged at 14,000 g for 5 min. The obtained pellets were immediately re-suspended in RNAlater solution (ThermoFisher Scientific, Waltham, MA USA) and stored at -80$_o$C until further processing. RNA extraction was performed with a previously described organic extraction protocol [43]. All RNA samples were treated with DNase using Turbo DNA-free kit (ambion, Austin, TX, USA), followed by removal of rRNA with Ribo-Zero rRNA removal kit (Epicentre, Madison, WI, USA). The enriched mRNA preparations were linearly amplified using the MessageAmp II-Bacteria Kit (Ambion), reverse transcribed with random hexamers using the Universal RiboClone cDNA Synthesis System (Promega, Madison, WI), followed by purification with the MiElute DNA clean-up kit (Qiagen). A total of four cDNA and 5 DNA libraries were prepared following the standard Illumina library preparation protocols, and sequenced on the Illumina GA II sequencer (paired end, 2×100; 300-500bp long inserts).

Table 5.1:Metagenomic libraries used for binning and genome recovery.

Dataset	Sample type	Time of sampling	Electron acceptor	Dataset Size (Gbp)
S3a	Sediment	2013	-	4.02
S3b	Sediment	2014	-	5.80
M7-PCE	Mesocosms	3 months after end of dechl.	PCE	3.8
M10-cDCE	Mesocosms	3 months after end of dechl.	cis-DCE	4.2
M5-TCE	Mesocosms	At the end of dechlorination	TCE	3.7
E3	**Sediment free enrichment**	At the end of dechlorination	TCE	4.2

The description of the above mesocosms and their phenotypic similarities and differences was reported previously [40]. In this study we used the above datasets for metagenomic binning in order to obtain high quality genomes from the E3 enrichment culture and hypothesize on the functional role of each consortium member.

5.3.4 Metagenomic binning and recovery of population genomes

For all DNA datasets, sequence reads were trimmed using a Q=15 Phred threshold and 50 bp minimum length after trimming. For the E3 metagenomic dataset, assembly was performed using IDBA-UB with default parameters [44]. For all other DNA datasets only the unassembled reads were used for metagenomic binning. Population genomes (bins) were recovered from the assembly of the E3 dataset using a combination of binning methods. First, the assembled contigs from the E3 dataset were binned with MaxBin [45], using coverage information from all obtained metagenomes. In brief, MaxBin calculates the tetra nucleotide frequency and the coverage (abundance) of each contig among the various datasets. Contigs are then binned together based on its high correlation in the abundance across datasets ($R^2 > 0.85$) or high correlation in tetra-nucleotide composition (distribution of tetra-nucleotides as Z-scores, $R^2 > 0.9$). Multiple metagenomic samples that contain the same organisms provide adequate data for robust correlations, and indeed binning was significantly improved as more metagenomes were used in this study (data not shown). The quality of the metagenomic bins was evaluated with CheckM [46], which employs the identification of single copy marker genes to estimate the completeness (fraction of the total marker genes recovered) and contamination (frequency of multiple copies of the marker genes recovered) for individual genomes.

Second, we used the phylogenetic information recovered in contigs longer than 5 Kbp in order to improve the recovered bins and eliminate any obvious contamination. Genes were predicted with GeneMark [47] and were taxonomically annotated using MyTaxa [48]. Contigs with >5 genes were inspected for consensus of taxonomic affiliations of the encoded genes, and contigs were reassigned or manually excluded accordingly. For example, several contigs from the *Dhc* bins remained un-binned when using MaxBin only, but taxonomic affiliations identified >3 genes nearly identical (>98% nucleotide identity) to reference *Dhc* genomes. We later realized that binning of the *Dhc* genome was probably affected by the presence of multiple *Dhc* strains within the E3 enrichment. Thus, we used both the changes in co-abundance (MaxBin) and the taxonomic information (Mytaxa) in order to curate the recovered genomes.

Lastly, an iterative assembly process was followed in order to improve the assembly quality of the obtained genomes: Reads from the E3 enrichment were mapped against the binned contigs using BLAT [49], and reads with >97% nucleotide identity and >98% alignment length were isolated for each bin, and reassembled in isolation from the remaining metagenomic dataset using the SPAdes assembler [50]. The newly obtained contigs were re-evaluated with CheckM and refined based on contig taxonomic classifications (MyTaxa step), and the re-assemby procedure was repeated until no further improvement in the N50 metric of the contigs length was observed.

Using this procedure, 14 bins were obtained with minimum contamination and high completeness levels. The final genome bins, and the set of all the contigs that remained un-binned were functionally annotated using the RAST genomic server [51]. Potentially complete pathways were identified using MinPath [52] and the obtained KEGG Orthology annotations [53] and manually curated based on the BRITE reference hierarchies, excluding the human diseases and organismal systems categories. Specific pathways of interest were evaluated by searching for the presence of characteristic enzymes identified from the literature and from the KEGG metabolic pathway maps. The list of genes that were evaluated for each pathway is provided in Table D (Appendix D).

5.3.5 Taxonomic characterization of metagenomes and recovered genomes

Taxonomic distributions in sequenced mesocosms and enrichments, were assessed based on identification of 16S rRNA reads with Parallel-META (v.2.1) [54], and classification against the SILVA database, release 111 [55]. Classified 16S reads were subsequently used for closed reference OTU picking (>97% identity threshold) with MacQIIME 1.8.0 [56]. OTU richness was estimated from read counts using the Chao1 index [57] with 95% confidence intervals as implemented in Chao1.pl from the enveomics collection [58]. OTU diversity was estimated using the Shannon index with correction for unobserved taxa [59] as implemented in the R package entropy [60]. OTU evenness was estimated as the ratio between the natural logarithm of estimated richness (Chao1) and the Shannon index (with Chao-Shen correction), an index typically referred to as Pielou's index.

The obtained genomes from the E3 enrichment culture were taxonomically characterized through phylogenetic reconstruction of their *rpoB* genes, which included

reference *rpoB* sequences form the closest relatives available in the NCBI genome collection. The *rpoB* genes from the bins and reference genomes were identified with with HMMer searches [61] using HMMER3 (http://hmmer.janelia.org/) with default settings and against the available Pfam models [62]. The identified RpoB amino acid sequences were aligned with Clustal-Omega [63], and the phylogenetic reconstruction was built with maximum likelihood using RAxML v7.4.2 [64] and 1,000 bootstraps, and the PROTGAMMAAUTO function to identify the best amino acid substitution model.

5.3.6 Metatranscriptome dataset processing and calculation of expression values

cDNA sequences were trimmed using the same criteria described in Chapter 2. In short, cDNA datasets were quality-filtered for probability of error cutoff of 1% at both ends, minimum length of 50 bp after trimming, and removal of N-containing sequences, poly-A tails, chimeric sequences, and non-coding RNA sequences (rRNAs and tRNAs). Only coupled reads that passed the above quality criteria were maintained for further analysis.

In order to obtain relative expression values of individual genes, the normalized coverage of the cDNA datasets was estimated as follows: cDNA reads were mapped against the all assembled genes (from both the recovered genomes and un-binned contigs) with BLAT and ≥98% alignment identity in ≥90% of the read length, and the BLAT results were filtered to include the best matches only. Read counts were estimated for each gene, and normalized for gene length and the total size of the cDNA dataset. The reported values are expressed on a per-gene basis as Reads Per Kbp Per Million reads (RPKM) unless otherwise noted. When the expression levels of individual pathways were assessed, RPKM values of the individual genes of the pathway were summed up to represent the specific pathway. Due to the lack of metagenomic datasets from the 4 time points examined, the abundance of each member at the DNA level couldn't be estimated. Thus no normalization was used for the abundance of the genomes in the enrichment, thus the RPKM values reflect relative transcriptional activity and abundance contribution at the total community level.

5.3.7 Identification and characterization of *RDase* genes and transcripts

RDase gene sequences were identified in the genomic bins and unassembled contigs by homology searches against a comprehensive *RDase* database (version as of 13-Oct-2016), which is continuously updated with newly characterized *RDases* from various sources [19]. Blastp search was performed and results were filtered with cutoffs for bit-score >80, alignment length >80% of the query sequence and >60% amino acid identity. Relative expression values for the identified *RDases* were obtained in a two step process. First, *RDase* reads were identified in the transcriptomes, using the ROCker pipeline [65] in order to minimize false positive matches of short reads against conserved regions shared with non- *RDases* proteins. Second, the identified *RDase* short reads were mapped back against the collection of *RDases* from the TCE enrichment in a competitive BLASTn search, and only best matches with >98% sequence nucleotide sequence were used.

5.3.8 Characterization of gene content diversity of the *Dhc* population

The high sequencing coverage of the E3 metagenomic sample allowed us to explore the intra-population diversity of the *Dhc* population using the GCD predictor, which is described in Chapter 4. Reads from the E3 metagenome were mapped against the recovered *Dhc* bin, and only best matches were retained. The GCD predictor identifies and quantifies genes within a genome that exhibit significantly lower coverage than the estimated core genome's average as a consequence of the variable distribution of genes within the *Dhc* populations. In other words, genes with lower coverage are present in a subset of the *Dhc* cells captured by the metagenomic bin (flexible), while genes with sequencing coverage close to the core genome's average are universally present in all the cells. Briefly, the statistical significance in the coverage difference is assessed by modeling the observed coverage of all genes (log-transformed) as a mix distribution resulting from combining an arbitrary number of skewed-normal distributions. Next, the skewed-normal distribution most likely corresponding to the core genome is selected; *i.e.*, that with highest average explaining more than 10% of the observations/genes. Finally, for each gene, the area under the core skewed-normal distribution below the observed coverage represents the p-value for the alternative hypothesis of a gene having a coverage significantly below the core. Genes with p-value below 0.05 are identified as putatively flexible.

5.3 RESULTS AND DISCUSSION

5.3.1 Taxonomic diversity of dechlorinating mesocosms

Three mesocosms were established with sediments isolated from a contaminated site within the Third Creek tributary (Fig 5.2), a location with a well reported history of localized contamination with a mixture of chlorinated ethenes [38]. All three mesocosms reductively dechlorinated the added chlorinated ethenes (PCE, TCE or cis-DCE) to ethene, presumably due to the presence of already adapted microbial communities to the compounds [40]. Metagenomic sequencing revealed that *Dhc* was among the most predominant members in all mesocosms, with abundance levels varying from 2 up to 35% of the total bacterial community, typically depending on the time at which the mesocosms was sampled (Fig 5.3). For example, *Dhc* were less abundant in mesocosms M7 and M10, for which DNA sampling was performed several days after the dechlorination had been completed, while reached up to 35% of the total community in the mesocosm M5, which was immediately sampled after the consumption of the last chlorinated ethene. Similarly, the abundance of *Dhc* was ~40% of the total bacterial cells in the sediment-free TCE dechlorinating culture E3.

As expected for enriched communities, the estimated bacterial diversity was substantially decreased from the original sediments in the slurry mesocosms, and was further decreased in the E3 culture (Table 5.2). For example, the contaminated site contained an estimated average of ~6,000 OTUs, with ~4,000 OTUs detected in the mesocosms, and only 2,500 of the estimated OTUs remained in the enriched community E3. Comparative metagenomic analysis of the mesocosms and detailed characterization of the dechlorination activity are described in Simsir et al [40]. Here, we used this collection of metagenomic datasets in order to identify the key community members that get established in the enriched sediment-free community E3, under the expectation that the most abundant members of the TCE enrichment culture would be present and relative abundant in all mesocosms. Indeed, the detection of core members with different relative abundances across samples (Fig 5.3) allowed for robust estimations of contig abundance co-variation, which was used for high-quality binning (see below).

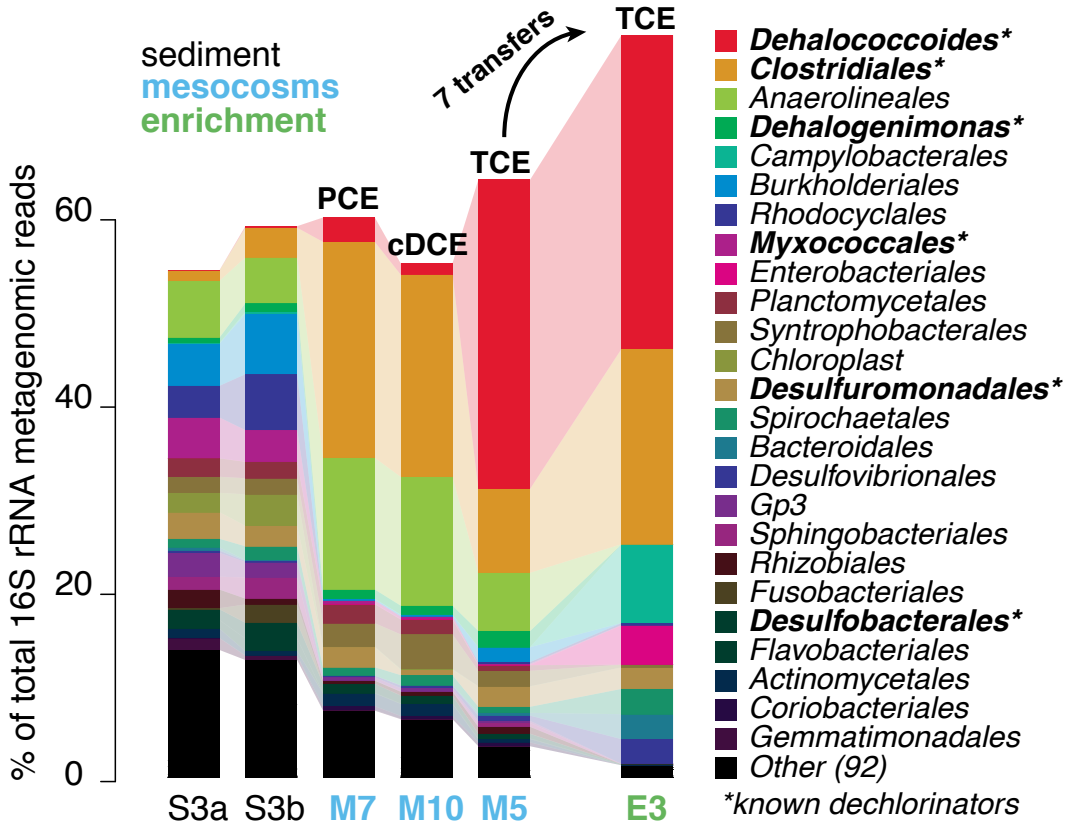


Figure 5.3: 16S rRNA gene-based taxonomic distributions of microbial communities in contaminated sediments and established dechlorinating mesocosms.

16S rRNA reads were identified in all available metagenomes and the taxonomic distributions are presented at the order level, except for *Dehalococcoidales* in which the two detected genera (*Dehalococcoides* and *Dehalogenimonas*) are shown separately. *Bacterial orders with known dechlorinator species.

Table 5.2: Bacterial diversity estimated on contaminated sediments and established enrichments.

Sample	Observed	Chao1	Chao1 $LB_{95\%}$	Chao1 $UB_{95\%}$	Shannon	Pielou
M10	2482	3942.1	3733.9	4184.9	6.2948	0.5712
M5	2152	3177.6	3021.2	3362.1	5.5672	0.6291
M7	2387	3808.9	3604.9	4046.9	6.5033	0.5506
S3-14	2054	5990.2	5712.8	6313.8	7.4161	0.4856
S3-12	3152	6196.1	5840.6	6598.6	7.9176	0.4789
E3-11	1505	2516.0	2326.1	2749.8	4.8716	0.6981

5.3.1 Characterization of the enriched TCE dechlorinating community

The obtained E3 enrichment culture was able to dechlorinate TCE to cis-DCE, VC and finally ethane, and no chlorinated substrates were detected after 77 days (Fig. 5.4A). Lactate was quickly consumed within the first five days, producing propionate and acetate, which gradually accumulated in the culture (Fig. 5.4B). *Dhc* cells, as monitored by 16S rRNA gene qPCR assays, increased in abundance during the incubation period, contributing to the increase in total bacterial 16S rRNA gene counts (Fig 5.4C). The highest bacterial and *Dhc* 16S rRNA gene counts were noted at the 54^{th} day, when the cis-DCE had been depleted from the culture, and were maintained in high proportions until the last day monitored, when the VC had been consumed. This observation is consistent with the expectation that Dhc are the only known organisms that can use cis-DCE and VC their growth under anaerobic conditions.

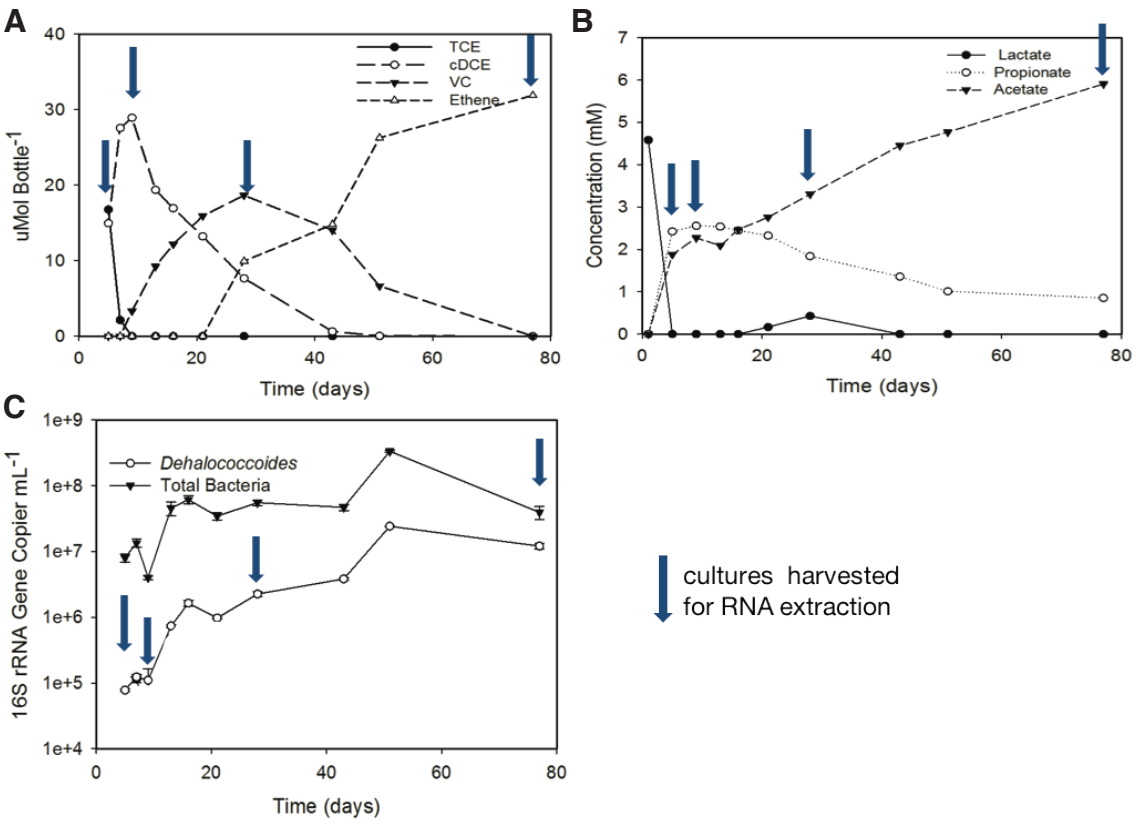


Figure 5.4: Time series monitoring of the TCE dechlorinating mixed community

(A) Concentrations of chlorinated ethenes and ethene, (B) lactate and fermentation byproducts, and (C) abundance of *Dhc* and total bacteria 16S rRNA genes were routinely monitored during the course of the dechlorination. Samples for RNA isolation and metatranscriptomic sequencing where taken at four time points, indicated by the blue arrows.

The successful dechlorination of TCE to ethene in this system implied that the four major requirements for mixed community dechlorination activity were met: (a) Dechlorinator strains capable of utilizing TCE, cis-DCE and VC were present in the enrichment, and it seemed likely that *Dhc* strains were the ones capable to catalyze the last two steps; (b) a series of fermentations and acetogenic reactions must have been taking place in order to convert lactate to hydrogen and acetate, both essential requirements for the Dhc metabolism [66,67]. (c) Finally *Dhc* relied on other community members for essential cofactors that cannot synthesize [21]. (d) The RDase enzymes and

dechlorinators were adequately protected from exposure to oxygen, which can otherwise lead to irreversible loss of dechlorination activity and kill or hamper anexic *Dhc* cultures [23,68]. We next aimed to identify community members responsible for the above requirements and their relative contributions to the efficiency of the dechlorination activity in this mixed community.

5.3.3 Recovery of abundant population genomes from the E3 culture

The E3 mixed community is dominated by nine major bacterial orders, each representing a different phylum, as evident by the 16S rRNA taxonomic distributions (Fig 5.3). *Dehalococcoidales*, *Clostridiales*, *Campylobacterales*, *Enterobacteriales*, *Desulfomonadales*, *Spirochaetales*, *Bacteroidales*, *Desulfovibrionales* and *Syntrophobacterales* representatives account for 82% of the total community based on the 16S taxonomic distributions. The low complexity E3 metagenome, in combination with the availability of metagenomic datasets from a series of microcosms allowed the recovery of 14 high-quality, manually curated bins (Table 5.3). The recovered genomes (bins) exhibited high levels of completeness and minimal contamination, and all together accounted for >85% of the total community DNA, and encoded up to 80% of the total community RNA (Fig. 5.5). The 14 identified genomes included members from all the major bacterial orders that were found to be abundant from the 16S rRNA analysis (Fig 5.3), with the exception of *Enterobacteriales*, for which high quality bins were not recovered, presumably due to the high complexity of this order in the metagenome.

Phylogenetic reconstructions of the identified *rpoB* genes from the bins confirmed that the recovered genome collection encompassed the majority of the microbial community in the E3 culture (Fig 5.4). In total, 16 *rpoB* genes were identified from the E3 assembled dataset, 14 of which were found in recovered genomes, and 2 were found in the remaining un-binned contigs (marked as UB in Fig 5.4). Thus, with the exception of an *Enterobacteriales* and a *Bacteroidetes* population, the recovered genomes encompassed all the major populations of the community.

Table 5.3: Recovered genomes from the E3 metagenomic dataset

Recovered Genome	ID	Taxonomy	Completeness (%)	Contamination (%)	Total length (bp)
Sphaerochata globosa E3	***Sph***	g__Sphaerochaeta	100	1.2	3,195,335
Sedimentobacter bacterium E3	***Sdm***	o__Clostridiales	99.49	5.1	4,688,194
***Lentimicrobium* bacterium E3**	***Ln1***	p__Bacteroidetes	99.46	1.61	5,190,274
Desulfovibrio desulfuricans E3	***Dsv***	g__Desulfovibrio	99.41	2.43	3,324,927
Arcobacter butzleri E3	***Arc***	g__Arcobacter	99.19	1.42	2,999,375
Sulfurospirilum bacterium E3	***Sul***	g__Sulfurospirillum	98.85	2.32	2,954,130
Clostridium bacterium E3	***Clo***	g__Clostridium	98.74	5.36	3,551,952
Desulfitobacterium bacterium E3	***Dsb***	f__Peptococcaceae	98.28	8.7	3,581,640
Acetobacterium bacterium E3	***Ace***	o__Clostridiales	98.12	6.57	2,360,669
Dehalococcoides mccartyi E3	***Dhc***	s__D._mccartyi	98.02	0	1,534,817
Geobacter lovleyi E3	***Geo***	g__Geobacter	96.91	1.94	3,327,696
***Lentimicrobium* bacterium2 E3**	***Ln2***	p__Bacteroidetes	96.77	1.93	3,920,709
***Bacteroideales* bacterium2 E3**	***Bac***	o__Bacteroidales	89.29	0.74	2,145,318
Deltaproteobacterium E3	***Del***	c__Deltaproteobacteria	86.97	2.76	2985906

Most of the recovered population genomes had been previously found in similar settings such as dechlorinating communities, or similar anaerobic systems where hydrogen is consumed. In the following section, we describe the 14 recovered members based on their abundance within the E3 culture .

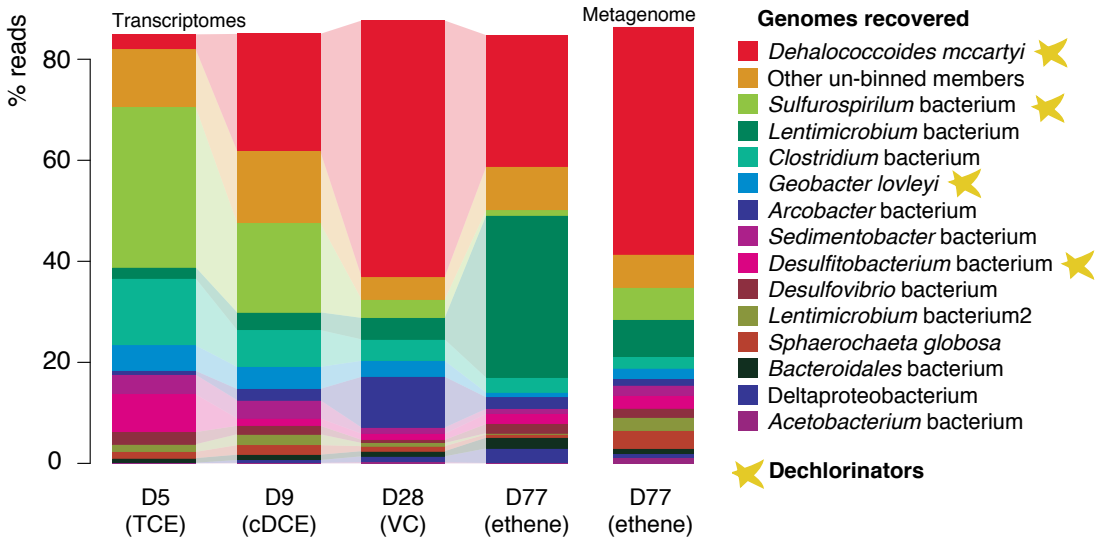


Figure 5.5: Whole-genome-based relative abundance of recovered genomes in the E3 culture.

Metagenomic and metatranscriptomic reads were mapped against the recovered genomes and un-binned contigs (category others; see figure key) to estimate their relative abundance, at the whole-genome level, in the datasets. Stars denote the identified dechlorinators in this system based on the presence of RDase sequences.

The recovered *Dhc* genome was classified within the only known species of ***Dehalococcoides mccartyi*** and as expected contained multiple RDase enzymes (see below). Four nearly complete genomes were obtained from the second most abundant order, ***Clostridiales,*** all of which are typically encountered in mixed dechlorinating communities with *Dhc* strains: (a) a *Sedimentobacter* population, nearly identical to the reference *Sedimentobacter* sp. B4 (99% Average Nucleotide Identity), an non-dechlorinating organism isolated from a reductively dechlorinating co-culture [69]; (b) a representative from the *Desulfitobacterium* genus, containing well characterized dechlorinating isolates that biodegrade a variety of chlorinated compounds [70]; (c) a genome classified as *Acetobacterium dehalogens*, a homoacetogenic bacterium that has been shown to grow with chlorinated methanes [71]; and (d) a genome of *Clostridum homopropionicum*, a strictly anaerobic organism that is capable of fermenting lactate to proprionate and acetate [72].

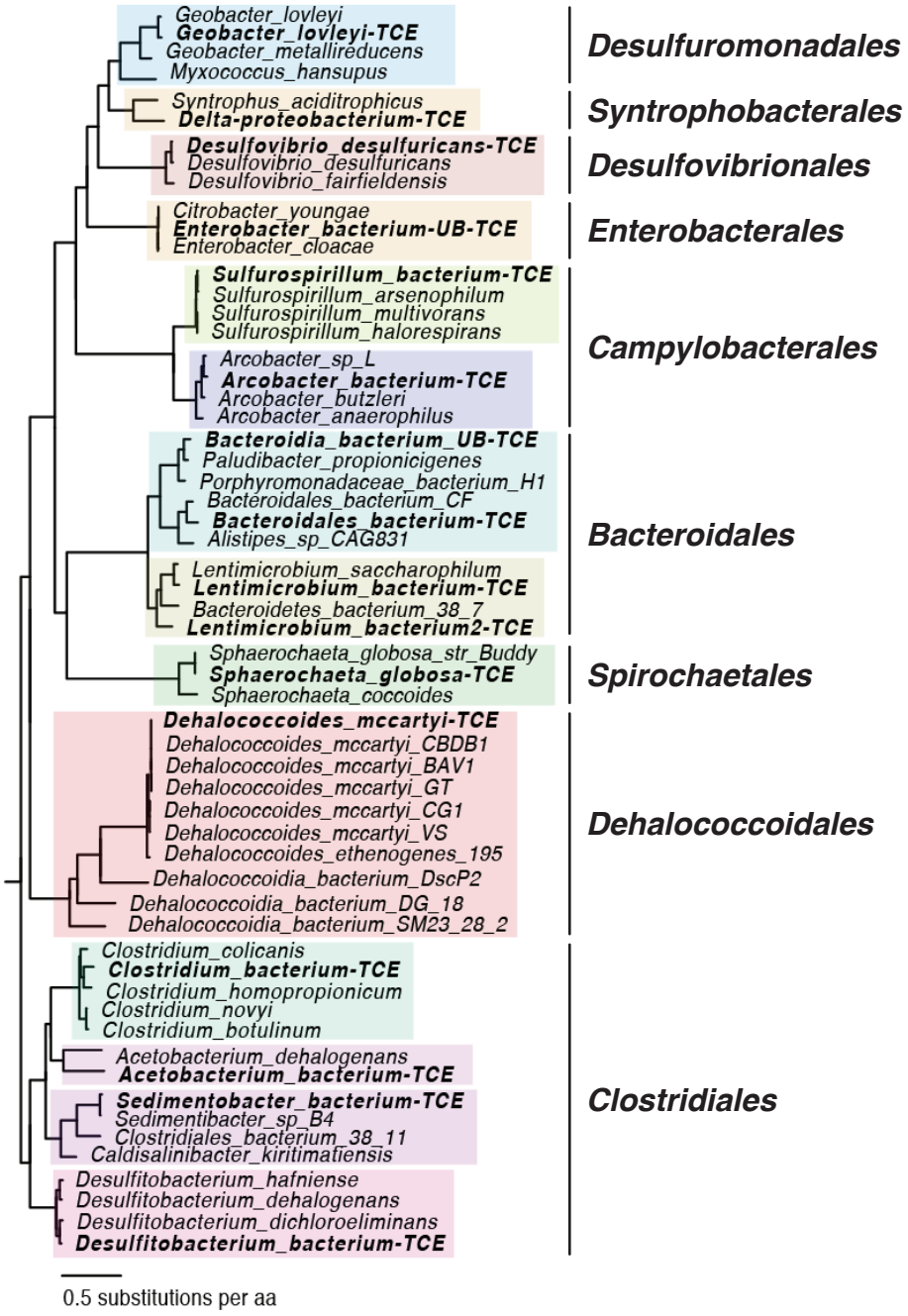


Figure 5.6: RpoB-based phylogenetic reconstruction of recovered genomes.

RpoB sequences from the E3 culture are shown in bold letters. The tree includes reference sequences from the closest relatives from the NCBI genome collection. Colors indicate classification within different bacterial genera. Note that, in addition to the 14 *rpoB* sequences identified within the recovered genomes, another two *rpoB* genes were found in the un-binned fraction of the assembled metagenome.

Two genomes were recovered that represented the ***Campylobacterales*** order: (a) A member of the *Sulfurospirilum* genus, a group with facultative anaerobic bacteria with versatile metabolism, including reductive dechlorination of PCE by some strains [73,74]; and (b) a member of the *Arcobacter* genus, which includes acetate-utilizing species commonly found in microbial fuel cell settings [75], but also within dechlorinating consortia, albeit less often [76]. The ***Desulfomonadales*** representative genome was classified within the *Geobacter lovleyi* species, a model dechlorinator organism that can grow on PCE and TCE dechlorination (among several other non chlorinated compounds), and is commonly found in *Dhc*-containing dechlorinating consortia [33,77]. Further, the recovered genomes included the fermentative *Sphaerochaeta globosa* (order ***Sphaerochaetales***), a species typically found in *Dhc*-containing enriched cultures when amended with lactate [78]. The ***Bacteroidales*** order included four of the recovered genomes. One of those members was found to be most similar to the reference genome of *Bacteroidales* bacterium CF, a non-dechlorinating, fermentative organism that was obtained from an enrichment culture that reductively dechlorinates chloroform or 1,1,1-trichloroethane [79]. The other two *Bacteroidales* genomes were classified within the *Lentimicrobium* genus, a taxon only recently described to contain strictly anaerobic bacteria that ferment a narrow range of carbohydrates, usually within anaerobic wastewater sludge systems [80]. The representative genome of ***Desulfovibrionales*** was classified within the *Desulfovibrio* genus, members of which are often found in synthrophic associations with *Dhc*, typically fermenting lactate to acetate and H_2 [27]. Finally, the genome representative of the ***Synthrophobacterales*** order, hereafter referred as *Deltaproteobacterium sp.*, was only remotely related to the *Syntrophus aciditrophicus* (54% AAI), a syntrophic bacterium found in anoxic settings along with H_2-consuming organisms [81].

5.3.4 Identification of dechlorinators in the mixed community

In order to identify the dechlorinator populations in the E3 consortium, we compared all assembled genes from both recovered genomes and un-binned contigs against a well curated RDase database containing sequences from the catalytic subunits of all known reductive dechlorinating enzymes that have been reported up to date [19]. Blastp-based searches identified a total of 51 nearly complete (average length 412 bp) putative RDase sequences (catalytic subunits), belonging to one of the four recovered

genomes (Table 5.4): *Dhc*, *Sulfurospirillum, Desulfitobacterium*, and *Geobacter*, all well characterized dechlorinators. Among the three organisms, only Dhc is an obligate dechlorinator and strictly hydrogenotrophic. In contrast, members of the *Sulfurospirillum*, *Geobacter* and *Desulfitobacterium* genera are versatile organisms, and can typically use a range of electron donors, such as organic acids and alcohols or hydrogen, and a range of electron acceptors such as fumarate, nitrate, various metals or some chlorinated compounds[82,83].

Five RDase genes were identified in the *Desulfitobacterium* genome, an unsurprising number, since up to seven RDases have been previously reported in dechlorinating *Desulfitobacter* isolates[83]. One RDase sequence was identified in the recovered *Geobacter* genome and agreement with this finding, the reference genomes *Geobacter lovleyi* encodes two putative RDase sequences [82]. Similarly, one RDase was identified for the *Sulfurospirillum* genome. Finally, 45 genes within the *Dhc* genome were identified as putative RDases, typically with high amino acid identities to their best-match reference RDase sequences, the latter encoded only by *Dhc* strains. An exception was noted for three RDase genes that were most closely related to the sequences from the *Dhc* strain BTF08[17] with ~42-61% amino acid identities, representing novel clades within the previously described diversity of dehalogenase enzymes. In most cases the identified Dhc RDases genes were found in the same contigs, adjacent to the genes that encode the membrane bound subunit of the enzyme, similarly to the RDase operon structures known from the available Dhc genomes [84].

Table 5.4: Identified RDase enzymes within the E3 TCE-dechlorinating enriched microbial consortium.

Identified *RDases*	Best match reference	Aminoacid identity	Reference strain
DhcTCE genome			
DhcTCE-UG34245*	cbdbA1595	92.9	*Dhc-cbdb*
DhcTCE-UG33339	8658278VS	51.39	*Dhc-VS*
DhcTCE-UG29148	AOV98866	61.28	*Dhc-Aov*
DhcTCE-UG27393	BAV1_0104	100	*Dhc-BAV*
DhcTCE-UG25226	BAV1_0112	100	*Dhc-BAV*
DhcTCE-UG23827*	AOV99976	95.92	*Dhc-Aov*

Table 5.4 continued

DhcTCE-UG23369	BAV1_0112	100	*Dhc-BAV*
DhcTCE-UG17812	BAV1_0104	99.64	*Dhc-BAV*
DhcTCE-G83	btf_1393	99.77	*Dhc-BTF*
DhcTCE-G598*	btf_1454	100	*Dhc-BTF*
DhcTCE-G594	cbdbA1539	100	*Dhc-cbdb*
DhcTCE-G591	cbdbA1542	100	*Dhc-cbdb*
DhcTCE-G587	AOV99962	100	*Dhc-Aov*
DhcTCE-G574	dcmb_1362	100	*Dhc-DCM5B*
DhcTCE-G546	cbdbA1508	85.84	*Dhc-cbdb*
DhcTCE-G495	dcmb_120	99.11	*Dhc-DCM5B*
DhcTCE-G449	DET1528	97.87	*Dhc-eth195*
DhcTCE-G446	JNA_RD19	100	*Dhc-JNA*
DhcTCE-G444	dcmb_1436	100	*Dhc-DCM5B*
DhcTCE-G443	dcmb_1434	98.99	*Dhc-DCM5B*
DhcTCE-G438	dcmb_1430	99.76	*Dhc-DCM5B*
DhcTCE-G436	dcmb_1428	99.47	*Dhc-DCM5B*
DhcTCE-G418	8658239VS	95.51	*Dhc-VS*
DhcTCE-G1563**	BAV1_0847	99.22	*Dhc-BAV*
DhcTCE-G1558	DET0876	78.98	*Dhc-eth195*
DhcTCE-G1546	GY50_1376	62.5	*Dhc-BTF08*
DhcTCE-G1480	BAV1_0173	100	*Dhc-BAV*
DhcTCE-G1478	cbdbA1560	100	*Dhc-cbdb*
DhcTCE-G1221	cbdbA1627	100	*Dhc-cbdb*
DhcTCE-G1213	cbdbA1638	99.6	*Dhc-cbdb*
DhcTCE-G1171	btf_1420	42.41	*Dhc-BTF*
DhcTCE-G1169	GT_1237	98.65	*Dhc-GT*
DhcTCE-G1153	dcmb_1370	100	*Dhc-DMC5*
DhcTCE-G1147	cbdbA1495	100	*Dhc-cbdb*
DhcTCE-G1143	KB13241_7	100	*Dhc-KB1*
DhcTCE-G1139	btf_1440	100	*Dhc-BTF*
DhcTCE-G1080	BAV1_0121	99.79	*Dhc-BAV*
DhcTCE-G1043	btf_1449	99.38	*Dhc-BTF*
DhcTCE-G1042	KB13241_1	100	*Dhc-KB1*
DhcTCE-G103	AOV99621	99.71	*Dhc-Aov*

Table 5.4 continued

DhcTCE-UG397	BAV1_0276	91.78	*Dhc-BAV*
DhcTCE-UG2632	BAV1_0119	99.36	*Dhc-BAV*
DhcTCE-UG2631	8658274VS	85.9	*Dhc-VS*
***Desulfitobacterium* genome**			
DesTCE-UG32299	ACT3-rdh14	71.37	*Dehalobacter CF*
DesTCE-UG28352	AJ439608	97.63	*Dehalobacter restrictus*
DesTCE-G3099	ACT3-rdh03	71.33	*Dehalobacter* CF
DesTCE-G3096	Dhaf_0711	91.8	*Desulfitobacterium hafniense*
DesTCE-G3094	Dhaf_0713	92.14	*Desulfitobacterium hafniense*
Sulfurospirillum			
SulTCE-G733	AF022812	67.54	*Sulfurospirillum multivorans*
***Geobacter* genome**			
GeobTCE-G849	Glov_2870	97.86	*Geobacter lovleyi SZ*

**The identified RDase DhcTCE-G1563 from the *Dhc* genomes is almost identical (99% amino acid identity) to the reference RDase BAV1_0847 (*bvcA*), encoded by the *Dhc* BAV1 isolate. This is the only RDase found within our collection that has been experimentally confirmed to dechlorinate cDCE and VC to ethene [85].

The presence of 45 RDases within the same genome is a surprising result as a maximum of 36 RDases have been identified in a single *Dch* genome to date [17]. Given the high completeness and quality of the recovered *Dhc* genome, this observation prompt us to further explore the micro-diversity of the *Dhc* population in the E3 enrichment. Recruitment analysis of the E3 dataset metagenomic reads against the recovered *Dhc* population genome revealed extended genomic regions with lower than average coverage (Fig 5.4). This observation can be explained by the presence of at least two *Dhc* strains (or more) within the E3 enrichment with gene content differences that are demonstrated as regions of lower than average coverage: genes present in only one strain (e.g., a subset of the *Dhc* population) would exhibit lower coverage compared to genes present in all strains (or the average coverage of the genome). The genes with lower coverage (n=197 from 1,800 total genes) were significantly enriched in unknown

and hypothetical proteins (Fisher's exact test, P < 2.2e-16) and were typically residing within identified prophage genomic regions. Other dispensable genes included 8 RDases and a gene encoding an ornithine cyclodeaminase (arginine/proline biosynthesis), all of which have been previously known to differentiate *Dhc* strains [86]. Taken together, the above observations indicate the presence of at least two *Dhc* strains in the TCE enrichment that contain up to 8 different RDases. This gene content diversity within a highly enriched community is somewhat unexpected, and implies that the competition between the different *Dhc* strains was likely not strong enough for one sub-population to outcompete the other.

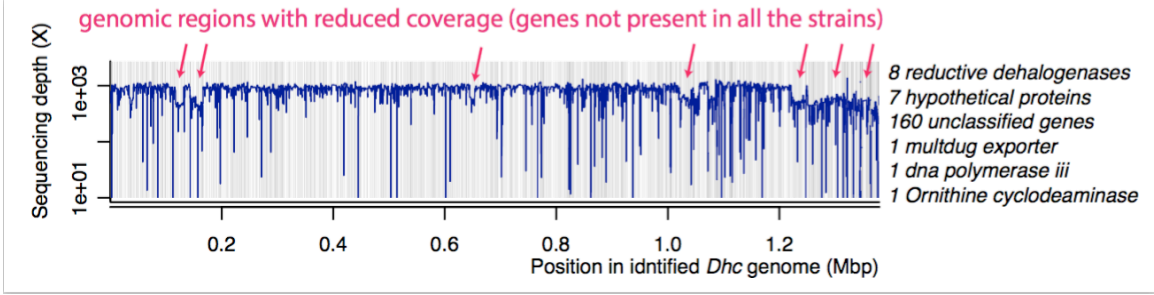


Figure 5.7: Read recruitment analysis for the identification of gene content differences in *Dhc* strains from the same sample.

High sequencing coverage allowed the assembly of a nearly complete *Dhc* genome, representing the consensus of the population present in the established TCE enrichment. When reads are recruited against the assembly, regions with lower coverage can be recognized as dispensable genes (i.e., present in some *Dhc* cells but not all).

Thus, the E3 enriched community contains at least four dechlorinator species, among which the *Dhc* population encompasses at least two strains with substantial gene content differences. No other RDase sequences were identified in the recovered bins or un-binned contigs. It is still possible that additional dechlorinators and genes are present in the enrichment and might have been missed from the assembled data. Such missed populations however must be relatively less abundant compared to populations recovered (rare biosphere) since at least 80% of the metagenomic reads were assembled into longer contigs for which open reading frames could be confidently predicted. Therefore, the identified RDases and dechlorinators represent the most abundant enzymes in the enriched consortium.

5.3.5 Transcriptional activity of RDases

The relative expression of the identified RDases revealed that all four identified dechlorinators had transcriptional activity of their RDase genes during the first time point, with the *Sulfurospirillum* being the most active when TCE was actively being consumed, while *Dhc* RDases dominated the transcript pool during subsequent times, with the highest transcript levels when the VC was the most abundant (Figure 5.8). The RDases from *Geobacter* and *Desulfitobacterium* were expressed in much lower levels during the first days, and were almost undetected at the last time points. The expression pattern of RDases reflected the overall transcriptional activity of the four dechlorinators. For example, the relative abundance of *Sulfurospirillum* transcripts in the cDNA datasets is maximum in the 5th day, when TCE is being consumed, and then it decreases during the other three time points. In contrast, the *Dhc* population gradually increases in abundance (16S rRNA gene counts) (Fig. 5.4), as does its relative transcriptional activity within the dechlorinating community (Fig. 5.5), with maximum abundance in the cDNA datasets during day 28, when the concentration of VC is at its maximum.

Gene expression levels of the individual RDases from the *Dhc* genome revealed that only three of the *Dhc* RDases accounted for the majority of RDase transcripts. Unsurprisingly, gene-1563 with 99% amino acid similarity with the reference BvcA RDase was among the most active gene (Figure 5.8). The BvcA RDase, is a typical biomarker for VC dechlorination, and one of the few enzymes that has been biochemically characterized and known to dechlorinate VC as well as cis-DCE [19]. The other two highly expressed RDases have close homologues in sequenced *Dhc* genomes, but haven't been functionally characterized. Their high expression levels in the transcriptome, in parallel with the growth and high abundance of *Dhc* in the E3 system, suggest these two enzymes might be able to dechlorinate cis-DCE and VC.

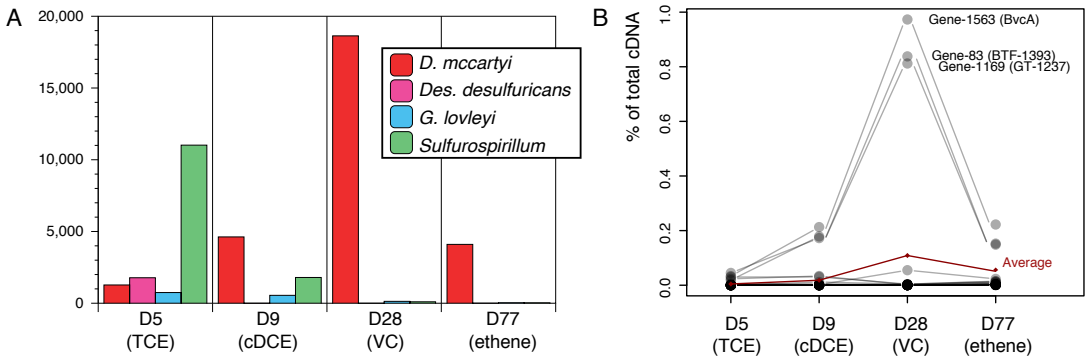


Figure 5.8: Transcriptional levels of RDase genes during the course of dechlorination.

A. Collective relative abundance of RDase transcripts from the four dechlorinators of the TCE enrichment expressed in RPKM values. B. Relative abundance of RDase transcripts for the *Dhc* enzymes through the course of dechlorination.

Taken together, our results showed that the major dechlorinators in this system apparently were Dhc, *Sulfurospirillum*, *Geobacter* and *Desulfitobacterium*. Based on the sequence identity of the recovered RDases against known references as well as their transcriptional profiles, it is most likely that the *Dhc* RDases are responsible for the cis-DCE and VC dechlorination, while the other three dechlorinators are participating in the TCE transformation. Interestingly, RDase transcripts from *Geobacter* were also detected during the second time point, right after the transformation of TCE, albeit at much lower levels. Additionally, the highest transcription of RDases was detected in the *Sulfurospirillum* genome, in accordance with the fact that several members of the genus are well characterize dechlorinators of TCE, and the transcriptional activity from this genome was at each maximum at the first time. However, *Sulfurospirillum* species are versatile respirators, thus the persistence of this organism at later time points during the incubation might be due to consumption of alterative electron acceptors in the system (ie fumarate) [74].

5.3.6 Positive interactions between *Dhc* and non dechlorinating community members

Consistent with the observations from all the available *Dhc* genomes up to date, the obtained *Dhc* from the E3 mesocosm lacked complete pathway biosynthetic pathways for at least three essential cofactors, namely corrinoids, biotin and thiamine. Thus, the accompanying community members were expected to contribute those essential nutrients directly or indirectly (i.e., as decaying biomass), since external vitamins were not provided with the media. In order to identify the community members that can provide essential nutrients to *Dhc*, we assessed the transcriptional profiles of 4 major metabolic pathways, which are described in detailed below.

Lactate fermentation

The first requirement for Dhc growth is the conversion of lactate to H_2 and acetate. Lactate can be fermented through various pathways, most of which start with the conversion to pyruvate through the lactate dehydrogenase enzymes. Alternative fermentation pathways can proceed to the production of acetate, ethanol and proprionate, and multiple pathways can proceed in parallel in the same organisms. For example several bacteria can ferment lactate to propionate and acetate either through the acryloyl-CoA or the methylmalonyl-CoA pathway[87]. Finally, acetogens can oxidize the propionate to acetate, CO_2 and H_2, a process thermodynamically favorable if an H_2-consuming reaction such as the dechlorination, is taking place [88].

Most of the non-dechlorinating genomes in the E3 culture encoded genes with the potential for lactate fermentation, typically towards acetate and H_2 directly (Fig 5.9). Surprisingly, we couldn't identify any genomes with the complete acryloyl-CoA pathway, through which lactate fermentation produces acetate, H_2 and formate, even thought *Clostridium* species are known to carry out this pathway [72]. It is possible that those genes might have been missed in the sequence assembly since only two genes are specific for this pathway. On the contrary, we identified most of the genes for the methylmalonyl-CoA pathway for propionate production in the two *Lentimicrobia* members.

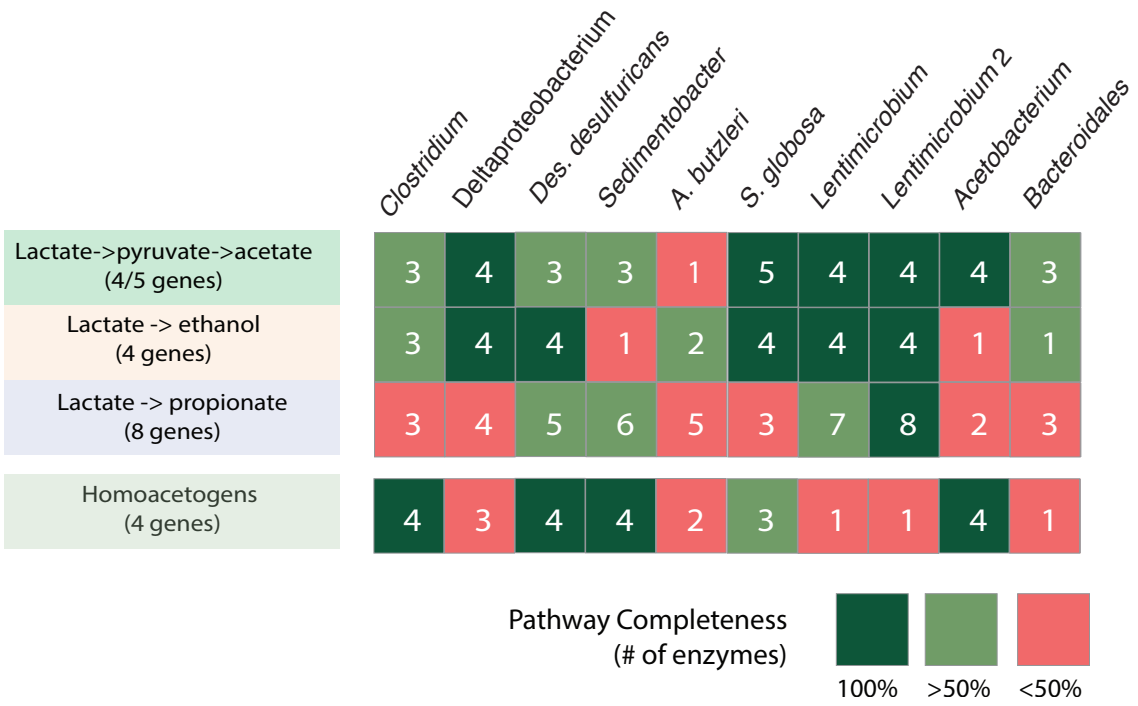


Figure 5.9: Major fermentation and acetogenesis pathways among the recovered genomes.

The lists of genes that were examined for each pathway are provided in the Table D1 (Appendix D). Pathways were identified based on the functional annotation of the genes encoded by the recovered population genomes using the available KEGG pathways. The completeness of each pathway is indicated by the color scale ranging from green to red, and the numbers of identified genes are reported for each genome.

Dechlorinators, methanogens or homoacetogens can consume the produced hydrogen from the fermentations [28]. No methanogenesis genes were identified among the recovered genomes (i.e., methyl coenzyme M reductase A, *mcrA*). The absence of methanogens was not surprising since those enrichments have been amended with high concentrations of the slowly fermentable lactate, which leads to low H_2 concentrations that ultimately select against methanogens. On the other hand, the genomic potential for homoacetogenesis was identified in at least four of the community members, including

the *Clostridium*, *D. desulfuricans*, *Sedimentobacter* and *Acetobacterium* members. While those organisms directly compete with *Dhc* for hydrogen, they might also have positive effects other than producing acetate, such as production of essential cofactors and vitamins.

Overall our predictions agreed with what has been previously known for several of the species present. For example *Desulfovibrio* is known to ferment lactate to acetate and hydrogen directly[27]. Among the identified H_2 scavengers, *Acetobacterium* representatives are known to perform homoacetogenesis[71].

Biotin biosynthesis

Biotin (or B7) is an essential cofactor for enzymes in lipid biosynthesis pathways, and growth of axenic cultures of *Dhc* requires external addition of biotin [21]. The *de novo* pathway involves two steps: the biosynthesis of the primelate moiety and the assembly of the bicyclic rings[89]. Various enzymes carry out the different reactions and often differ among organisms. However, the assembly of bicyclic rings is an evolutionary conserved pathway that is carried out by 4 enzymes (BioF, A, D, B) [89]. The operon encoding the four enzymes was found complete or nearly complete for five members among the recovered community genomes (Fig. 5.10). Among them *Sulfurosprillum* and *G. lovleyi* exhibited the highest transcriptional activity during the first days (5 and 79), and *A. bultzeri* during day 28, in accordance to their total transcriptional activity (Fig. 5.6).

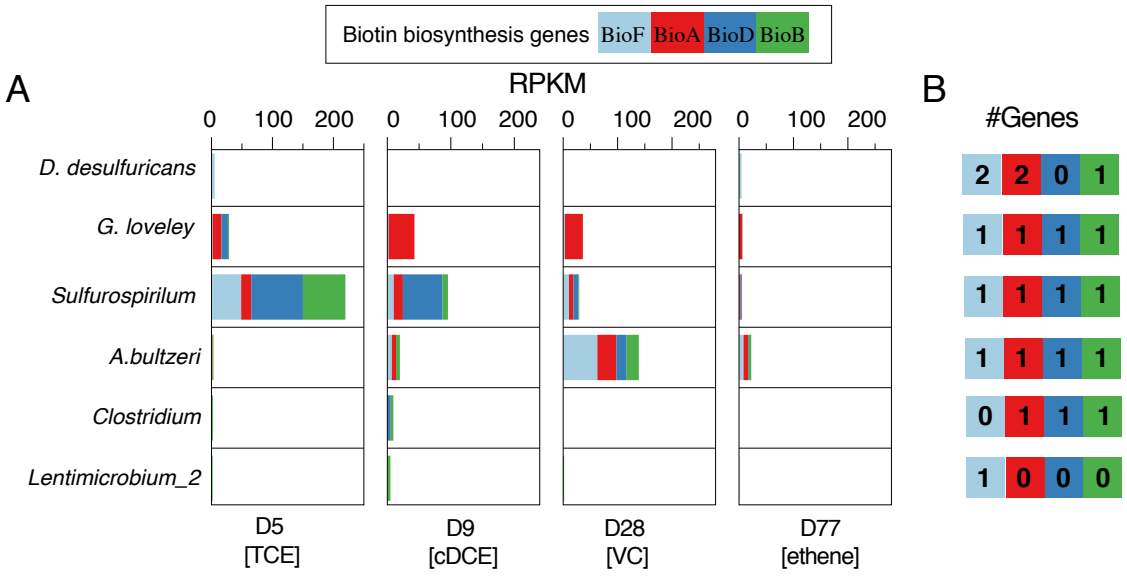


Figure 5.10: Relative expression of biotin biosynthesis genes.

(A) Expression values for the four biotin biosynthetic pathways during the course of dechlorination. (B) Four enzymes of the biotin synthesis pathway (assembly of bicyclic rings) are required (among others) and are conserved among species (top); the graph represents the number of biotin biosynthetic genes identified for each step of the pathway in the recovered genomes.

Thiamine biosynthesis

Thiamine (or vitamin B_1) is an essential cofactor required for the activity of several anabolic enzymes [90]. Thiamine diphosphate is the active form, and there are at least three enzymes required for its production: hydroxymethyl pyrimidine synthase (ThiC), thiamine phosphate synthase (ThiE) and hydroxymethyl pyrimidine (phosphate) kinase (ThiD), all of which were found in 9 out of 14 recovered genomes (Fig 10.11). Transcriptional activity of the aforementioned enzymes was most prominent for *Sulfurospirillum* and *G. lovleyi* at the onset of dechlorination (day 5), as well as for *A. butzeri* and the Deltaproteobacterium during the day 28. However, transcripts were detected for all the genomes encoding those genes, albeit at much lower levels for most members.

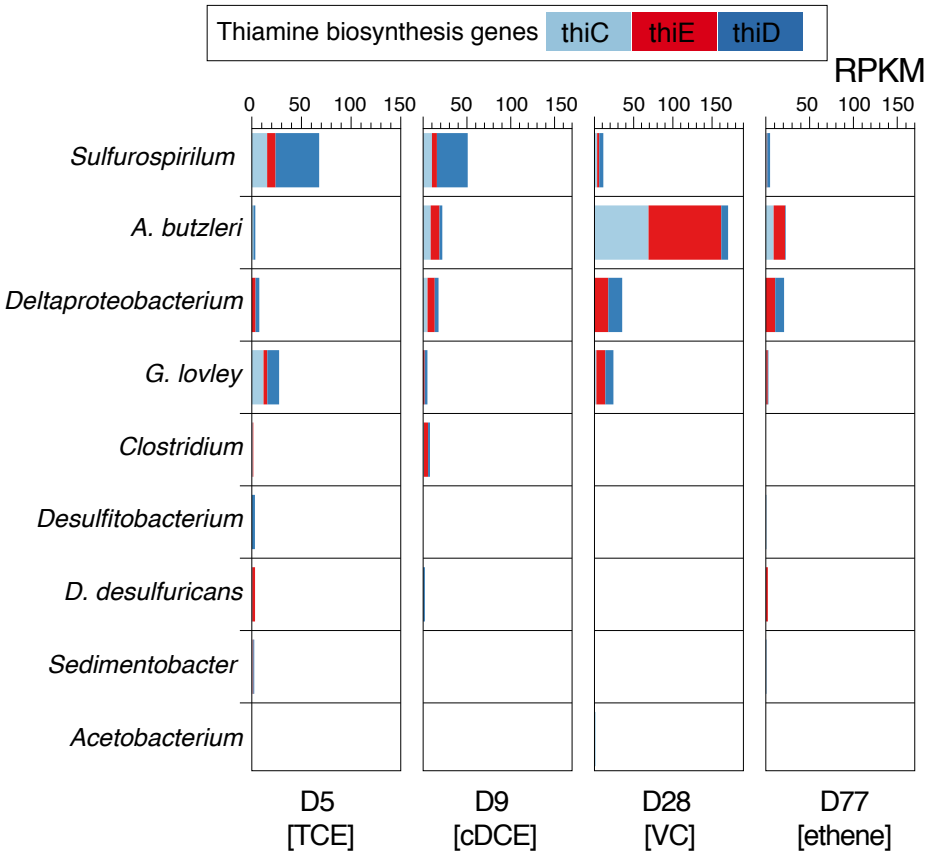


Figure 5.11: Relative expression of thiamine biosynthesis genes.
Expression values for three if the required biosynthetic enzymes for de novo thiamine synthesis. Only the genomes for which the three examined thiamine biosynthesis genes were identified are shown.

Corrinoid biosynthesis

The complete biosynthetic pathway for corrinoid biosynthesis is lacking from the recovered *Dhc* genome of the E3 culture, as well as from all the *Dhc* genomes reported to-date, despite the fact that cobalamin, a prominent member of the corrinoid group, is the essential cofactor of RDase enzymes [12,21]. Instead, the *Dhc* genomes, including the recovered *Dhc* from the E3 culture, encode multiple pathways for importing and remodeling various corrinoids, which can be subsequently transformed to adenosylcobalamin, the active form of vitamin B_{12} (Fig 5.12A)[91]. Within the E3 consortium, all recovered genomes encoded corrinoid transporters, and various pathways for remodeling the imported corrinoids, such as cobyrinate, cobinamide, and

the lower ligand DMB (5,6-dimethylbenzimidazole), as well as direct uptake of cobalamin (Fig 5.12B). However, only seven genomes were found to contain most of the genes required for the *de novo* corrinoid biosynthesis, i.e., conversion of precorrin-2 to cobyrinate. *Sulfurospirillum*, *Clostridium* and *Desulfitobacterium* genomes encoded all the 10 enzymes required, while the Deltaproteobacterium, *Geobacter*, *D. desulfuricans* and *Sedimentibacter* contained at least seven of them (some genes could have been missed due to sequencing gaps and/or misassemblies). Among those genes, *cbiK* was found in all seven genomes with genomic evidence of a complete biosynthetic pathway, thus this gene was used as a proxy for the relative expression of the *de novo* corrinoid synthesis (Fig 5.12).

The most predominant member in the transcription of the *de novo* corrinoid biosynthesis pathway at the onset of dechlorination (day 5) was *Sulfurospirillum*, along with relatively smaller contributions from *Geobacter*, *D. desulfuricans* and Deltaproteobacterium genomes. Active biosynthesis and remodeling of corrinoids was expected for the dechlorinators *Geobacter* and *D. desulfuricans*, in accordance with transcriptional expression of their RDases during the time that TCE was being consumed. Interestingly, the expression of corrinoid biosynthesis pathway was not detected during the subsequent days, even at the time when the *Dhc* cells were the most active (i.e., day 28, during the VC utilization phase). This observation might be explained by the fact that the produced corrinoids from the highly active *Sulfurospirillum* genome during the first days provided adequate supply of the cofactor for *Dhc* to complete the dechlorination.

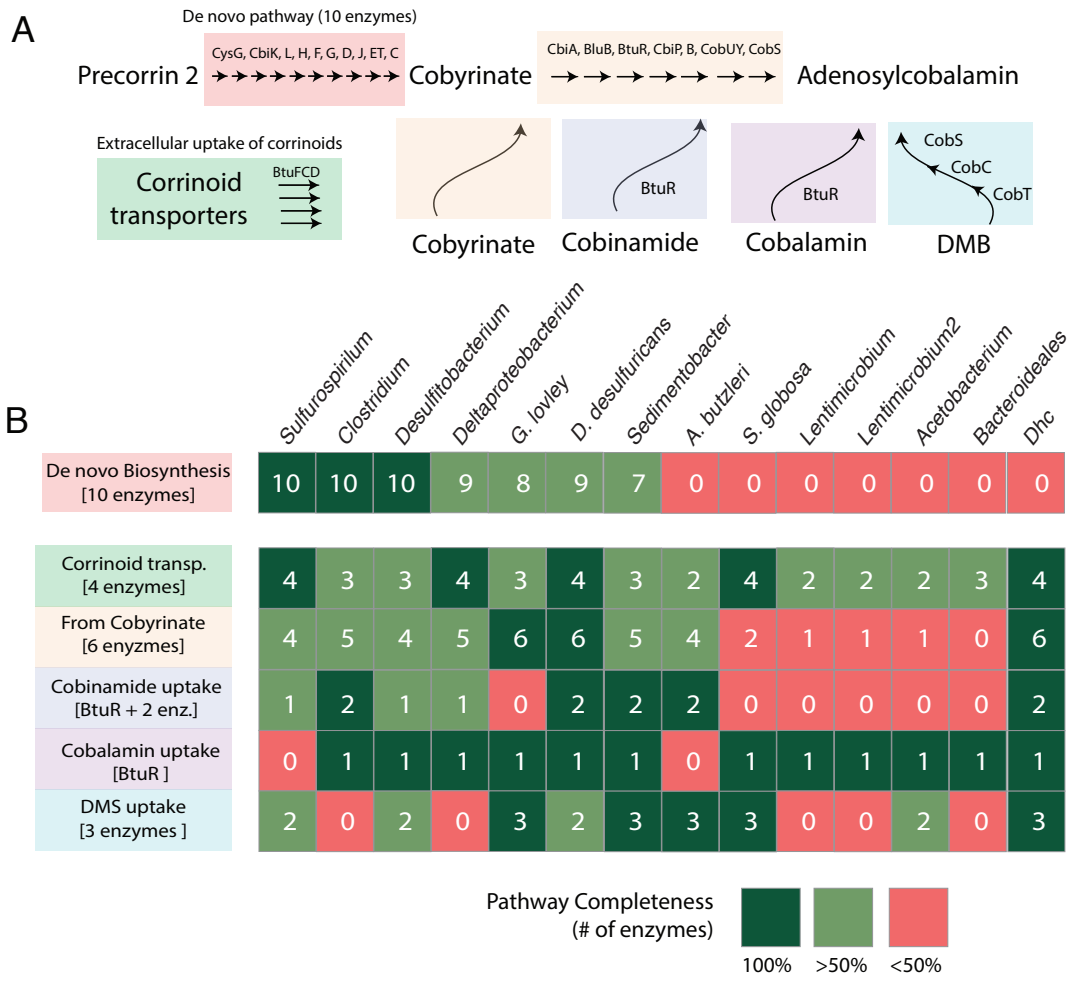


Figure 5.12: Cobalamin biosynthesis and scavenging pathways in the 14 recovered genomes of the enrichment.

(A) The first steps of the *de novo* biosynthetic pathway involve 10 enzymes, which can transform precorrin-2 to cobyrinate. Another 7 enzymes are required to transform cobyrinate to the active form adenosylcobalamin. In addition, corrinoid transporters can be used to import various corrinoids that can eventually be converted to adenosylcobalamin [31,91,92]. (B) Presence of *de novo* biosynthesis and scavenging of corrinoids in the members of the E3 consortium. The completeness of each pathway is indicated by the color scale ranging from green to red, and the numbers of identified genes are reported for each genome.

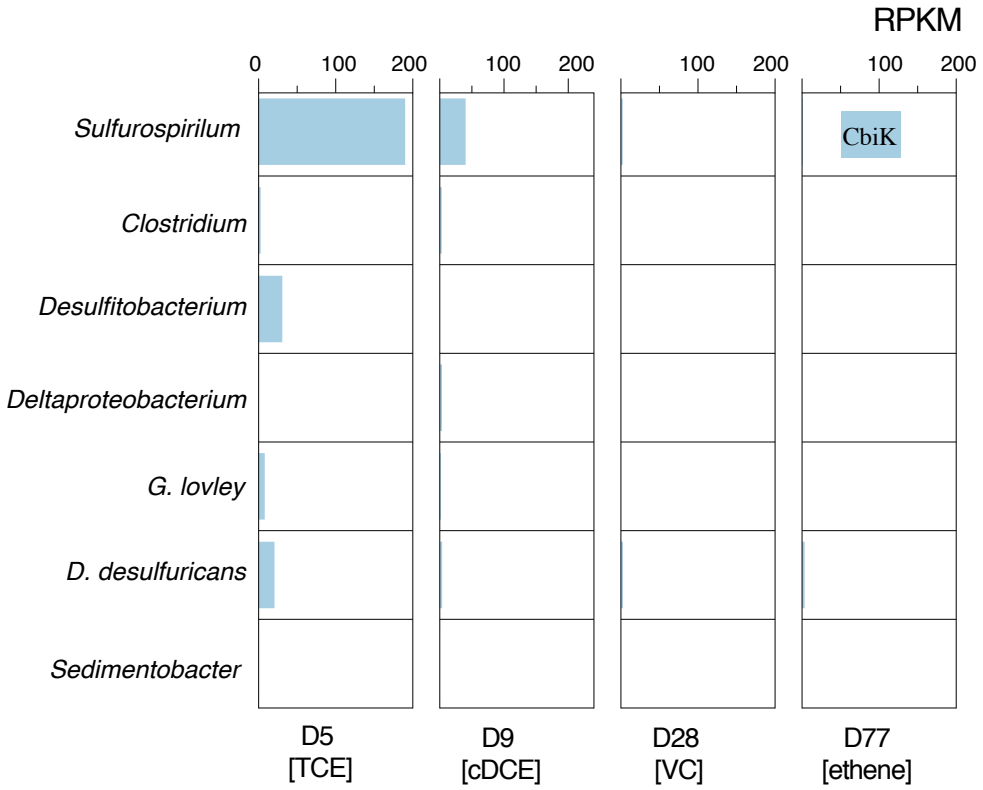


Figure 5.13: Expression of *de novo* cobalamin biosynthesis pathway during dechlorination.

Only genomes that encoded the complete biosynthesis pathway are shown. The CbiK enzyme was used as an indicator for the relative expression of the *de novo* branch of the cobalamin biosynthesis pathway.

Oxygen scavenging

Maintenance of anaerobic conditions is critical for the sustainability of the dechlorination reactions [23]. It has been hypothesized that the presence of *Dhc* within mixed communities results in more robust cultures partially due to the presence of multiple oxygen scavenging pathways that are encoded by other community members [31]. The recovered Dhc genome from the E3 mesocosm encodes at least 6 genes that are putatively involved in the protection from radical oxygen species: one superoxide dismutase (SOD), one catalase, two rubrerythrin and two rubredoxin enzymes (Figure 5.14, Table D1). All four enzymes encoded by Dhc are transcriptionally active at all time

points examined, with maximum expression values when *Dhc* is most active and abundant (day 28), and at the end of dechlorination (day 77).

Additional enzymes for the protection from radical oxygen species [93], or the direct utilization of oxygen (oxidases and peroxidases) are encoded in all community members. Among them, at least ten members showed active transcription at some time during the incubation (Fig 5.14). For example, at day 5, during the onset of the TCE dechlorination, the genomes with the most significant contributions in expression of oxygen scavenging enzymes were *Sulfurospirillum* and *Geobacter*, as might have been expected since those two organisms dominated the total transcriptome at this time point. *Dhc* was less dominant in the transcriptomes at the 5th day, but collectively expressed more RDase transcripts compared to the other two dechlorinators. As the dechlorination proceeded and *Dhc* became more abundant during the last two time points, several other members showed significant contribution to oxidative stress response transcripts, including the *Lentimicrobium*, *A. bultzeri* and *D. desulfiricans* genomes. While the expression patterns of those enzymes were highly dynamic, and the major contributors varied across the time points, collectively the community members expressed more than 90% of the total transcripts related to oxygen scavenging, an observation that could explain the significantly improved sustainability of mixed versus axenic *Dhc* cultures.

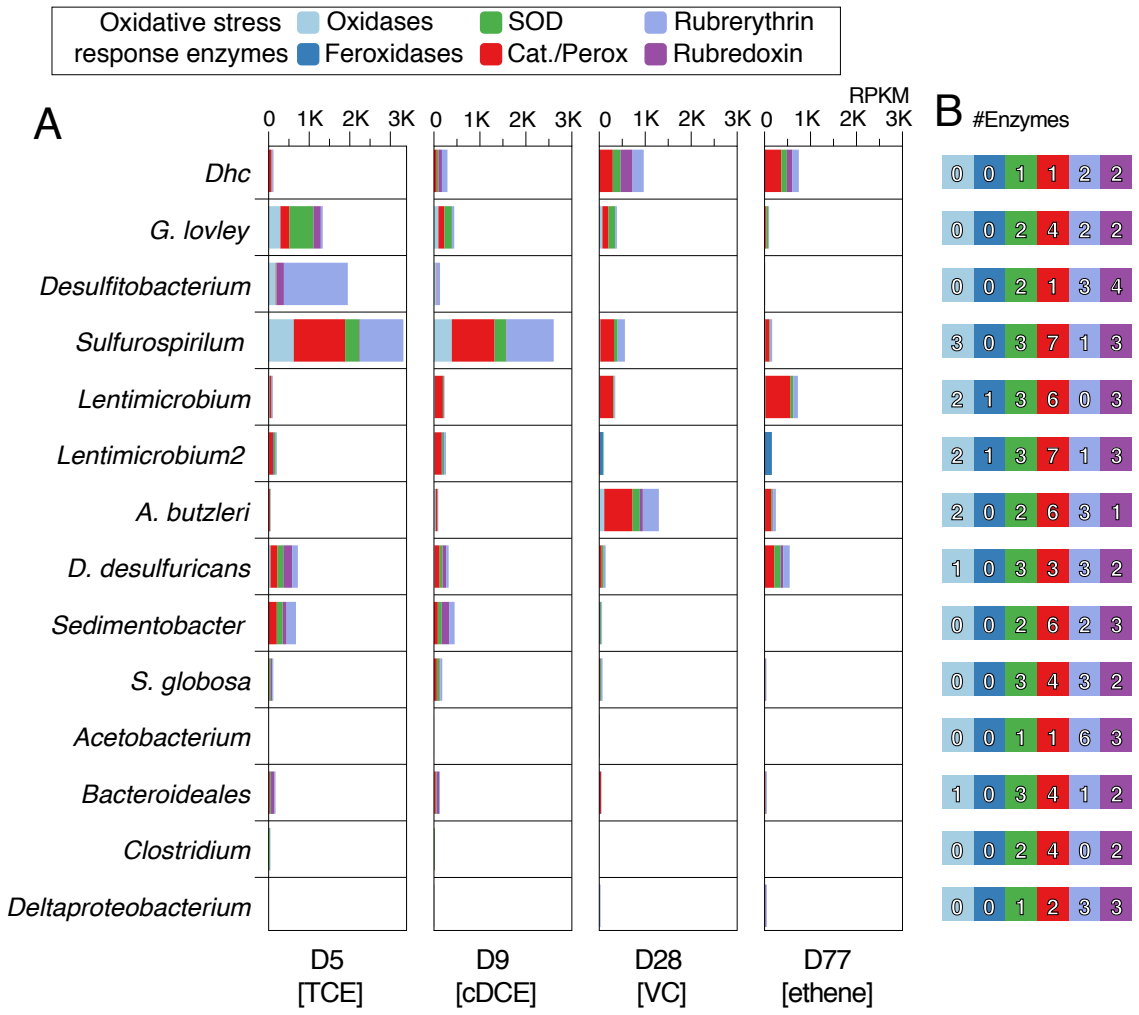


Figure 5.14: Oxidative stress response enzymes and their expression patterns during the course of dechlorination.

(A) Expression values for oxidative stress response enzymes for each genome are provided as the added RPKM values of all identified genes related to oxygen scavenging and protection from reactive oxygen species. (B) Number of identified genes for each category of oxidative stress response enzymes are provided for each genome. The list of the enzymes examined is provided in Table D1.

5.5 CONCLUSIONS

The TCE-dechlorinating community examined in this study has been maintained for the past 2 years, through repeated transfers in fresh medium and amendment with TCE and lactate. The resulting microbial consortium is dominated by few highly enriched populations, which have been continuously adapting to the same conditions. In this highly enriched community, 14 microbial members made up ~85% of the total bacteria, as revealed by quantitative metagenomic analysis. Among them, four dechlorinator species exhibited various contributions in the dechlorination activity throughout time: All four dechlorinators *Sulfurospirillum, Geobacter lovleyi*, *Desulfitobacterium* and Dhc actively expressed RDases during the TCE dechlorination, but *Dhc* dominated the RDase transcript pool during subsequent time points when VC and cis-DCE was being consumed. We identified three highly expressed *Dhc* RDase genes throughout the course of dechlorination, two of which haven't been linked before to VC and cDCE dechlorination activity, and can thus, be target of future functional verification experiments and used as potential biomarkers for the evaluation of *in situ* dechlorination.

The non-dechlorinating community members encoded overlapping metabolic potentials and their relative transcriptional contributions were dynamic at the different times examined. The availability of nearly complete genomes for those members enabled us to assign potential key roles and assess their relative contribution towards the *Dhc* nutritional requirements, based on gene transcript relative abundance. The non dechlorinating community members were predicted to synergistically produce acetate and hydrogen through fermentation and acetogenic reactions, while a smaller subset, and predominately the dechlorinators, was responsible for *de novo* biosynthesis of essential vitamins, including B12, thiamine and biotin. In particular, the *Sulfurospirillum*, and, to a lesser extend, the *Arcobacter* and Deltaproteobacterium members appeared to be responsible for the large majority of *de novo* biosynthesis of biotin, thiamine and B_{12} precursors (Fig 5.15). Additionally, the community transcriptome showed that non-dechlorinating community members expressed high levels of enzymes involved in oxygen scavenging, in accordance with previous hypothesis for their role in the maintenance of the culture robustness, especially when compared to axenic *Dhc* cultures.

Finally, despite the long term enrichment of the E3 community, *Dhc* strains with different gene repertoires were maintained within the same culture, presumably due the absence of intrapopulation competition. For example, at least eight RDase genes were found to be different among *Dhc* strains, but none of those enzymes was identified as highly expressed in this system. Future experimentation with the E3 culture under different conditions might reveal the functional role of those flexible genes.

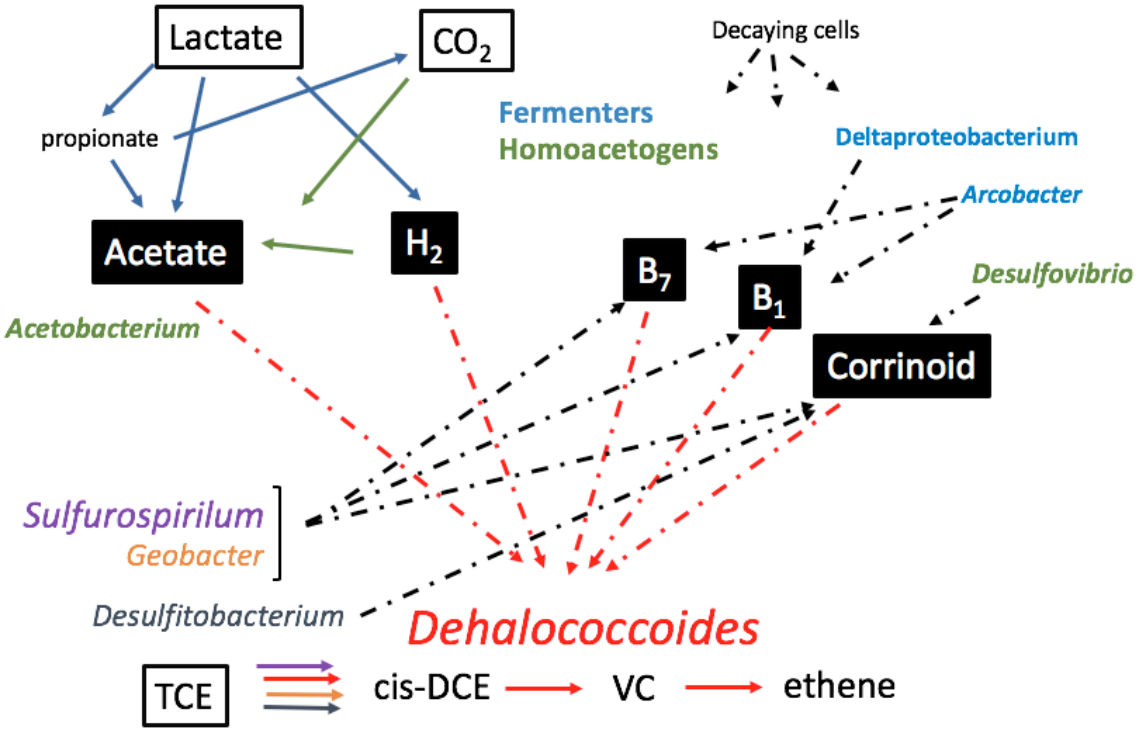


Figure 5.15: Schematic representation of the metabolic process likely affecting the Dhc population within the TCE dechlorinating mixed consortium.

Multiple community members were predicted to be responsible for the initial transformation of lactate to hydrogen and acetate, which were subsequently used from Dhc as electron donor and carbon source, based on gene expression patterns. Dhc was predicted to be the major dechlorinator during the last steps of the process, while *Geobacter* and *Desulfitobacterium* are potential contributors in the TCE dechlorination. Additionally, the *Sulfurospirillum*, *Geobacter*, *Arcobacter* and *Deltaproteobacterium* members were the major contributors of the de novo synthesized essential vitamins required by Dhc.

Taken together, our results provided insights into the dynamic microbial interactions of the mixed community during active dechlorination, and identified potential supportive mechanisms for the robust activity and growth of the fastidious *Dhc* population. The putative identified interactions from the transcriptomic data can guide future experimentation and assessment of the proposed mechanisms presented here, to evaluate the specific contributions of the key community members identified. Nevertheless, our results revealed that among the various community members dechlorinators might be the major contributors of vitamins required by Dhc. Thus the presence of polychlorinated ethenes and growth of TCE dechlorinators might facilitate the Dhc growth and the subsequent complete dechlorination of less chlorinated ethenes such as cDCE and VC.

5.6 REFERENCES

1. Gribble, G. W. The diversity of naturally produced organohalogens. *Chemosphere* **52,** 289–297 (2003).
2. Doherty, R. E. A History of the Production and Use of Carbon Tetrachloride, Tetrachloroethylene, Trichloroethylene and 1,1,1-Trichloroethane in the United States: Part 1—Historical Background; Carbon Tetrachloride and Tetrachloroethylene. *Environ. Forensics* **1,** 69–81 (2000).
3. Holliger, C., Wohlfarth, G. & Diekert, G. Reductive dechlorination in the energy metabolism of anaerobic bacteria. *FEMS Microbiol. Rev.* **22,** 383–398 (1998).
4. Pant, P. & Pant, S. A review: advances in microbial remediation of trichloroethylene (TCE). *J. Environ. Sci. China* **22,** 116–126 (2010).
5. Maphosa, F. *et al.* Ecogenomics of microbial communities in bioremediation of chlorinated contaminated sites. *Front. Microbiol.* **3,** (2012).
6. Jugder, B.-E. *et al.* Organohalide Respiring Bacteria and Reductive Dehalogenases: Key Tools in Organohalide Bioremediation. *Front. Microbiol.* **7,** 249 (2016).
7. Chambon, J. C. *et al.* Review of reactive kinetic models describing reductive dechlorination of chlorinated ethenes in soil and groundwater. *Biotechnol. Bioeng.* **110,** 1–23 (2013).
8. He, J., Ritalahti, K. M., Aiello, M. R. & Loffler, F. E. Complete Detoxification of Vinyl Chloride by an Anaerobic Enrichment Culture and Identification of the Reductively Dechlorinating Population as a Dehalococcoides Species. *Appl. Environ. Microbiol.* **69,** 996–1003 (2003).
9. He, J., Sung, Y., Krajmalnik-Brown, R., Ritalahti, K. M. & Löffler, F. E. Isolation and characterization of Dehalococcoides sp. strain FL2, a trichloroethene (TCE)- and 1,2-dichloroethene-respiring anaerobe. *Environ. Microbiol.* **7,** 1442–1450 (2005).
10. Lee, P. K. H., Cheng, D., West, K. A., Alvarez-Cohen, L. & He, J. Isolation of two new Dehalococcoides mccartyi strains with dissimilar dechlorination functions and

their characterization by comparative genomics via microarray analysis. *Environ. Microbiol.* **15,** 2293–2305 (2013).
11. Taş, N., Van Eekert, M. H. A., De Vos, W. M. & Smidt, H. The little bacteria that can – diversity, genomics and ecophysiology of 'Dehalococcoides' spp. in contaminated environments. *Microb. Biotechnol.* **3,** 389–402 (2010).
12. Löffler, F. E. *et al.* Dehalococcoides mccartyi gen. nov., sp. nov., obligately organohalide-respiring anaerobic bacteria relevant to halogen cycling and bioremediation, belong to a novel bacterial class, Dehalococcoidia classis nov., order Dehalococcoidales ord. nov. and family Dehalococcoidaceae fam. nov., within the phylum Chloroflexi. *Int. J. Syst. Evol. Microbiol.* **63,** 625–635 (2013).
13. Richardson, R. E. Genomic insights into organohalide respiration. *Curr. Opin. Biotechnol.* **24,** 498–505 (2013).
14. Adrian, L., Hansen, S. K., Fung, J. M., Görisch, H. & Zinder, S. H. Growth of Dehalococcoides strains with chlorophenols as electron acceptors. *Environ. Sci. Technol.* **41,** 2318–2323 (2007).
15. Adrian, L., Szewzyk, U., Wecke, J. & Görisch, H. Bacterial dehalorespiration with chlorinated benzenes. *Nature* **408,** 580–583 (2000).
16. Cupples, A. M., Spormann, A. M. & McCarty, P. L. Growth of a Dehalococcoides-like microorganism on vinyl chloride and cis-dichloroethene as electron acceptors as determined by competitive PCR. *Appl. Environ. Microbiol.* **69,** 953–959 (2003).
17. Pöritz, M. *et al.* Genome sequences of two dehalogenation specialists – Dehalococcoides mccartyi strains BTF08 and DCMB5 enriched from the highly polluted Bitterfeld region. *FEMS Microbiol. Lett.* **343,** 101–104 (2013).
18. Uchino, Y. *et al.* Complete genome sequencing of Dehalococcoides sp. strain UCH007 using a differential reads picking method. *Stand. Genomic Sci.* **10,** 102 (2015).
19. Hug, L. A. *et al.* Overview of organohalide-respiring bacteria and a proposal for a classification system for reductive dehalogenases. *Phil Trans R Soc B* **368,** 20120322 (2013).
20. Ahsanul Islam, M., Edwards, E. A. & Mahadevan, R. Characterizing the Metabolism of Dehalococcoides with a Constraint-Based Model. *PLoS Comput Biol* **6,** e1000887 (2010).
21. Schipp, C. J., Marco-Urrea, E., Kublik, A., Seifert, J. & Adrian, L. Organic cofactors in the metabolism of Dehalococcoides mccartyi strains. *Philos. Trans. R. Soc. B Biol. Sci.* **368,** (2013).
22. Vainberg, S., Condee, C. W. & Steffan, R. J. Large-scale production of bacterial consortia for remediation of chlorinated solvent-contaminated groundwater. *J. Ind. Microbiol. Biotechnol.* **36,** 1189–1197 (2009).
23. Maymó-Gatell, X., Chien, Y., Gossett, J. M. & Zinder, S. H. Isolation of a Bacterium That Reductively Dechlorinates Tetrachloroethene to Ethene. *Science* **276,** 1568–1571 (1997).
24. Duhamel, M. & Edwards, E. A. Microbial composition of chlorinated ethene-degrading cultures dominated by Dehalococcoides. *FEMS Microbiol. Ecol.* **58,** 538–549 (2006).
25. He, J., Holmes, V. F., Lee, P. K. H. & Alvarez-Cohen, L. Influence of Vitamin B12 and Cocultures on the Growth of Dehalococcoides Isolates in Defined Medium. *Appl. Environ. Microbiol.* **73,** 2847–2853 (2007).

26. Cheng, D. & He, J. Isolation and characterization of 'Dehalococcoides' sp. strain MB, which dechlorinates tetrachloroethene to trans-1,2-dichloroethene. *Appl. Environ. Microbiol.* **75,** 5910–5918 (2009).
27. Men, Y. *et al.* Sustainable syntrophic growth of Dehalococcoides ethenogenes strain 195 with Desulfovibrio vulgaris Hildenborough and Methanobacterium congolense: global transcriptomic and proteomic analyses. *ISME J.* **6,** 410–421 (2012).
28. Duhamel, M. & Edwards, E. A. Growth and Yields of Dechlorinators, Acetogens, and Methanogens during Reductive Dechlorination of Chlorinated Ethenes and Dihaloelimination of 1,2-Dichloroethane. *Environ. Sci. Technol.* **41,** 2303–2310 (2007).
29. Freeborn, R. A. *et al.* Phylogenetic analysis of TCE-dechlorinating consortia enriched on a variety of electron donors. *Environ. Sci. Technol.* **39,** 8358–8368 (2005).
30. Richardson, R. E., Bhupathiraju, V. K., Song, D. L., Goulet, T. A. & Alvarez-Cohen, L. Phylogenetic characterization of microbial communities that reductively dechlorinate TCE based upon a combination of molecular techniques. *Environ. Sci. Technol.* **36,** 2652–2662 (2002).
31. Hug, L. A., Beiko, R. G., Rowe, A. R., Richardson, R. E. & Edwards, E. A. Comparative metagenomics of three Dehalococcoides-containing enrichment cultures: the role of the non-dechlorinating community. *BMC Genomics* **13,** 327 (2012).
32. Rowe, A. R., Lazar, B. J., Morris, R. M. & Richardson, R. E. Characterization of the community structure of a dechlorinating mixed culture and comparisons of gene expression in planktonic and biofloc-associated 'Dehalococcoides' and Methanospirillum species. *Appl. Environ. Microbiol.* **74,** 6709–6719 (2008).
33. Yan, J., Ritalahti, K. M., Wagner, D. D. & Löffler, F. E. Unexpected specificity of interspecies cobamide transfer from Geobacter spp. to organohalide-respiring Dehalococcoides mccartyi strains. *Appl. Environ. Microbiol.* **78,** 6630–6636 (2012).
34. Lai, Y. & Becker, J. G. Compounded effects of chlorinated ethene inhibition on ecological interactions and population abundance in a Dehalococcoides - Dehalobacter coculture. *Environ. Sci. Technol.* **47,** 1518–1525 (2013).
35. Yan, J., Im, J., Yang, Y. & Löffler, F. E. Guided cobalamin biosynthesis supports Dehalococcoides mccartyi reductive dechlorination activity. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **368,** 20120320 (2013).
36. Cheng, D., Chow, W. L. & He, J. A Dehalococcoides-containing co-culture that dechlorinates tetrachloroethene to trans-1,2-dichloroethene. *ISME J.* **4,** 88–97 (2010).
37. Lee, L. K., Ding, C., Yang, K.-L. & He, J. Complete debromination of tetra- and penta-brominated diphenyl ethers by a coculture consisting of dehalococcoides and desulfovibrio species. *Environ. Sci. Technol.* **45,** 8475–8482 (2011).
38. Şimşir, B., Yan, J., Graves, D. & Löffler, F. Natural Attenuation in Streambed Sediment Receiving Chlorinated Solvents from Underlying Fracture Networks.
39. Löffler, F. E., Sanford, R. A. & Ritalahti, K. M. Enrichment, cultivation, and detection of reductively dechlorinating bacteria. *Methods Enzymol.* **397,** 77–111 (2005).
40. Şimşir, B., Tsementzi, D., Konstantinidis, K. & Löffler, F. E. Comparative metagenomics of active dechlorinating mesocosms from a contamined creek and upstream pristine locations. *Prep.*

41. Amos, B. K., Christ, J. A., Abriola, L. M., Pennell, K. D. & Löffler, F. E. Experimental Evaluation and Mathematical Modeling of Microbially Enhanced Tetrachloroethene (PCE) Dissolution. *Environ. Sci. Technol.* **41,** 963–970 (2007).
42. Ritalahti, K. M. *et al.* Quantitative PCR Targeting 16S rRNA and Reductive Dehalogenase Genes Simultaneously Monitors Multiple Dehalococcoides Strains. *Appl. Environ. Microbiol.* **72,** 2765–2774 (2006).
43. Tsementzi, D., Poretsky, R., Rodriguez-R, L. M., Luo, C. & Konstantinidis, K. T. Evaluation of metatranscriptomic protocols and application to the study of freshwater microbial communities. *Environ. Microbiol. Rep.* **6,** 640–655 (2014).
44. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinforma. Oxf. Engl.* **28,** 1420–1428 (2012).
45. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32,** 605–607 (2016).
46. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25,** 1043–1055 (2015).
47. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26,** 1107–1115 (1998).
48. Luo, C., Rodriguez-R, L. M. & Konstantinidis, K. T. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Res.* **42,** e73 (2014).
49. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
50. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).
51. Aziz, R. K. *et al.* The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9,** 75 (2008).
52. Ye, Y. & Doak, T. G. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput. Biol.* **5,** e1000465 (2009).
53. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40,** D109-114 (2012).
54. Su, X., Xu, J. & Ning, K. Parallel-META: efficient metagenomic data analysis based on high-performance computation. *BMC Syst. Biol.* **6 Suppl 1,** S16 (2012).
55. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35,** 7188–7196 (2007).
56. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7,** 335–336 (2010).
57. Chao, A. Nonparametric Estimation of the Number of Classes in a Population. *Scand. J. Stat.* **11,** 265–270 (1984).
58. Rodriguez-R, L. M. & Konstantinidis, K. T. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. (2016). doi:10.7287/peerj.preprints.1900v1
59. Chao, A. & Shen, T.-J. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.* **10,** 429–443

60. Hausser, J. & Strimmer, K. Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks. *J. Mach. Learn. Res.* **10,** 1469--1484 (2009).
61. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7,** e1002195 (2011).
62. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42,** D222–D230 (2014).
63. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539–539 (2014).
64. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinforma. Oxf. Engl.* **22,** 2688–2690 (2006).
65. Orellana, L. H., Rodriguez-R, L. M. & Konstantinidis, K. T. ROCker: accurate detection and quantification of target genes in short-read metagenomic data sets by modeling sliding-window bitscores. *Nucleic Acids Res.* gkw900 (2016). doi:10.1093/nar/gkw900
66. Delgado, A. G., Parameswaran, P., Fajardo-Williams, D., Halden, R. U. & Krajmalnik-Brown, R. Role of bicarbonate as a pH buffer and electron sink in microbial dechlorination of chloroethenes. *Microb. Cell Factories* **11,** 128 (2012).
67. Ziv-El, M. *et al.* Using electron balances and molecular techniques to assess trichoroethene-induced shifts to a dechlorinating microbial community. *Biotechnol. Bioeng.* **109,** 2230–2239 (2012).
68. Amos, B. K., Ritalahti, K. M., Cruz-Garcia, C., Padilla-Crespo, E. & Löffler, F. E. Oxygen Effect on Dehalococcoides Viability and Biomarker Quantification. *Environ. Sci. Technol.* **42,** 5718–5726 (2008).
69. Maphosa, F., van Passel, M. W. J., de Vos, W. M. & Smidt, H. Metagenome analysis reveals yet unexplored reductive dechlorinating potential of Dehalobacter sp. E1 growing in co-culture with Sedimentibacter sp. *Environ. Microbiol. Rep.* **4,** 604–616 (2012).
70. Marzorati, M. *et al.* A Novel Reductive Dehalogenase, Identified in a Contaminated Groundwater Enrichment Culture and in Desulfitobacterium dichloroeliminans Strain DCA1, Is Linked to Dehalogenation of 1,2-Dichloroethane. *Appl. Environ. Microbiol.* **73,** 2990–2999 (2007).
71. Schilhabel, A. *et al.* The Ether-Cleaving Methyltransferase System of the Strict Anaerobe Acetobacterium dehalogenans: Analysis and Expression of the Encoding Genes. *J. Bacteriol.* **191,** 588–599 (2009).
72. Beck, M. H. *et al.* Draft Genome Sequence of the Strict Anaerobe Clostridium homopropionicum LuHBu1 (DSM 5847). *Genome Announc.* **3,** (2015).
73. Buttet, G. F., Holliger, C. & Maillard, J. Functional Genotyping of Sulfurospirillum spp. in Mixed Cultures Allowed the Identification of a New Tetrachloroethene Reductive Dehalogenase. *Appl. Environ. Microbiol.* **79,** 6941–6947 (2013).
74. Goris, T. *et al.* Insights into organohalide respiration and the versatile catabolism of Sulfurospirillum multivorans gained from comparative genomics and physiological studies. *Environ. Microbiol.* **16,** 3562–3580 (2014).
75. Toh, H. *et al.* Complete Genome Sequences of Arcobacter butzleri ED-1 and Arcobacter sp. Strain L, Both Isolated from a Microbial Fuel Cell. *J. Bacteriol.* **193,** 6411–6412 (2011).
76. redOrbit. Characterization of a Microbial Consortium Capable of Rapid and Simultaneous Dechlorination of 1,1,2,2-Tetrachloroethane and Chlorinated Ethane

and Ethene Intermediates. *Redorbit* (2007). Available at: http://www.redorbit.com/news/science/797119/characterization_of_a_microbial_consortium_capable_of_rapid_and_simultaneous/. (Accessed: 31st October 2016)
77. Sung, Y. *et al.* Geobacter lovleyi sp. nov. Strain SZ, a Novel Metal-Reducing and Tetrachloroethene-Dechlorinating Bacterium. *Appl. Environ. Microbiol.* **72,** 2775–2782 (2006).
78. Ritalahti, K. M. *et al.* Sphaerochaeta globosa gen. nov., sp. nov. and Sphaerochaeta pleomorpha sp. nov., free-living, spherical spirochaetes. *Int. J. Syst. Evol. Microbiol.* **62,** 210–216 (2012).
79. Tang, S. & Edwards, E. A. Complete Genome Sequence of Bacteroidales Strain CF from a Chloroform-Dechlorinating Enrichment Culture. *Genome Announc.* **1,** (2013).
80. Sun, L. *et al.* Lentimicrobium saccharophilum gen. nov., sp. nov., a strictly anaerobic bacterium representing a new family in the phylum Bacteroidetes, and proposal of Lentimicrobiaceae fam. nov. *Int. J. Syst. Evol. Microbiol.* **66,** 2635–2642 (2016).
81. McInerney, M. J. *et al.* The genome of Syntrophus aciditrophicus: Life at the thermodynamic limit of microbial growth. *Proc. Natl. Acad. Sci.* **104,** 7600–7605 (2007).
82. Wagner, D. D. *et al.* Genomic determinants of organohalide-respiration in Geobacter lovleyi, an unusual member of the Geobacteraceae. *BMC Genomics* **13,** 200 (2012).
83. Kim, S.-H. *et al.* Genome sequence of Desulfitobacterium hafnienseDCB-2, a Gram-positive anaerobe capable of dehalogenation and metal reduction. *BMC Microbiol.* **12,** 21 (2012).
84. McMurdie, P. J. *et al.* Localized Plasticity in the Streamlined Genomes of Vinyl Chloride Respiring Dehalococcoides. *PLoS Genet* **5,** e1000714 (2009).
85. Krajmalnik-Brown, R. *et al.* Genetic Identification of a Putative Vinyl Chloride Reductase in Dehalococcoides sp. Strain BAV1. *Appl. Environ. Microbiol.* **70,** 6347–6351 (2004).
86. Marco-Urrea, E., Seifert, J., Bergen, M. von & Adrian, L. Stable Isotope Peptide Mass Spectrometry To Decipher Amino Acid Metabolism in Dehalococcoides Strain CBDB1. *J. Bacteriol.* **194,** 4169–4177 (2012).
87. Seeliger, S., Janssen, P. H. & Schink, B. Energetics and kinetics of lactate fermentation to acetate and propionate via methylmalonyl-CoA or acrylyl-CoA. *FEMS Microbiol. Lett.* **211,** 65–70 (2002).
88. Azizian, M. F., Marshall, I. P. G., Behrens, S., Spormann, A. M. & Semprini, L. Comparison of lactate, formate, and propionate as hydrogen donors for the reductive dehalogenation of trichloroethene in a continuous-flow column. *J. Contam. Hydrol.* **113,** 77–92 (2010).
89. Lin, S. & Cronan, J. E. Closing in on complete pathways of biotin biosynthesis. *Mol. BioSyst.* **7,** 1811–1821 (2011).
90. Du, Q., Wang, H. & Xie, J. Thiamin (Vitamin B1) Biosynthesis and Regulation: A Rich Source of Antimicrobial Drug Targets? *Int. J. Biol. Sci.* **7,** 41–52 (2011).
91. Yi, S. *et al.* Versatility in Corrinoid Salvaging and Remodeling Pathways Supports Corrinoid-Dependent Metabolism in Dehalococcoides mccartyi. *Appl. Environ. Microbiol.* **78,** 7745–7752 (2012).
92. Rodionov, D. A., Vitreschak, A. G., Mironov, A. A. & Gelfand, M. S. Comparative Genomics of the Vitamin B12 Metabolism and Regulation in Prokaryotes. *J. Biol. Chem.* **278,** 41148–41159 (2003).
93. Imlay, J. A. The molecular mechanisms and physiological consequences of oxidative stress: lessons from a model bacterium. *Nat. Rev. Microbiol.* **11,** 443–454 (2013).

CHAPTER 6

SINGLE CELL GENOMES AND METAGENOMES LINK SAR11 BACTERIA WITH ANOXIA AND OCEAN NITROGEN LOSS

Reproduced in part with permission from D. Tsementzi, J. Wu, S. Deutsch, S. Nath, LM Rodriguez-R, A.S. Burns, P. Ranjan, N. Sarode, R. R. Malmstrom, C. C. Padilla, B. K. Stone, Laura A. Bristow, M. Larsen, J. B. Glass, B. Thamdrup, T. Woyke, K. T. Konstantinidis, F. J. Stewart. *Nature Microbiol.* 2016 Aug 16.

6.1 ABSTRACT

Bacteria of the SAR11 clade constitute up to one half of all microbial cells in the oxygen-rich surface ocean. DNA sequences from SAR11 are also abundant in oxygen minimum zones (OMZs) where oxygen falls below detection and anaerobic microbes play important roles in converting bioavailable nitrogen to N_2 gas. Evidence for anaerobic metabolism in SAR11 has not yet been observed, and the question of how these bacteria contribute to OMZ biogeochemical cycling is unanswered. Here, we identify the metabolic basis for SAR11 activity in anoxic ocean waters. Genomic analysis of single cells from the world's largest OMZ revealed diverse and previously uncharacterized SAR11 lineages that peak in abundance at anoxic depths, but are largely undetectable in oxygen-rich ocean regions. OMZ SAR11 contain adaptations to low oxygen, including genes for respiratory nitrate reductases (Nar). SAR11 *nar* genes were experimentally verified to encode proteins catalyzing the nitrite-producing first step of denitrification and constituted ~40% of all OMZ *nar* transcripts, with transcription peaking in the zone of maximum nitrate reduction rates. These results redefine the ecological niche of Earth's most abundant organismal group and suggest an important contribution of SAR11 to nitrite production in OMZs, and thus to pathways of ocean nitrogen loss.

6.2 INTRODUCTION

Alphaproteobacteria of the SAR11 clade form one of the most ecologically dominant organism groups on the planet, representing up to half of the total microbial community in the oxygen-rich surface ocean[1–5]. All characterized SAR11 isolates, including the globally ubiquitous *Pelagibacter* genus, are aerobic heterotrophs adapted for scavenging dissolved organic carbon and nutrients under the oligotrophic conditions of the open ocean[6–9]. Gene-based surveys have also revealed diverse SAR11 lineages at high abundance in the deep waters of the meso- and bathypelagic realms[10–13]. However, the functional properties that distinguish SAR11 living in distinct ocean regions remain unclear. All known SAR11 genomes are small (typically less than 1.5 Mbp), with genomic streamlining as a potential adaptation to the nutrient limiting conditions of the open ocean.[11] It has been hypothesized that adaptations in SAR11 do not involve large variations in gene content[6,8], suggesting that SAR11's contribution to ocean biogeochemistry is primarily through its role in aerobic oxidation of organic carbon. Although genetic or biochemical evidence of anaerobic metabolism has not been reported for SAR11, high abundances of SAR11-related genes have been detected under anoxic conditions in marine oxygen minimum zones (OMZs). Permanent OMZs extend over ~8% of the oceanic surface area (O_2 < 20 μM)[14], with the largest and most intense OMZs in upwelling regions of the Eastern Pacific. In the cores of these regions microbial respiration of high surface primary production combines with low ventilation to deplete oxygen (O_2) from mid-water depths, resulting in O_2 concentrations below detection (~10 nM) over a major portion (~100-700 m) of the water column[15]. In the absence of O_2, respiratory nitrate (NO_3^-) reduction to nitrite (NO_2^-) becomes the dominant process for organic matter oxidation [16], with respiratory NO_3^-reductases (Nar) being among the most abundant and highly expressed enzymes in OMZs [17–19]. NO_3^- respiration results in a substantial accumulation of NO_2^- in OMZs, often to micromolar concentrations[20]. This NO_2^- pool is actively cycled through NO_2^--consuming microbial metabolisms, notably the anaerobic processes of denitrification and anaerobic ammonium oxidation (anammox)[21,22], which together in OMZs account for 30-50% of the loss of bioavailable nitrogen from the ocean as either gaseous dinitrogen (N_2) or nitrous oxide (N_2O)[21,22]. Surprisingly, SAR11 bacteria are often the most abundant organisms in

the NO_2^--enriched N-loss zone of OMZs where O_2 is undetectable, representing ~20% (range: 10-40%) of all 16S rRNA genes and protein-coding metagenome sequences in the 0.2 to 1.6 μm biomass fraction [18,19,23,24]. Such high abundances imply that SAR11 make up a substantial fraction of the OMZ community and raise the question of SAR11's role in OMZ biogeochemistry.

Here, we analyzed single amplified genomes (SAG) to identify the metabolic basis for SAR11's dominance in anoxic OMZs. We focused on SAR11 SAGs obtained from the Eastern Tropical North Pacific (ETNP) OMZ off Mexico, the world's largest OMZ accounting for 41% of global OMZ surface area[14] (Fig. 6.1a). Oxygen concentration ($[O_2]$) at this site declined from ~200 µM at the surface to ~400 nM at the bottom of the oxycline (30-85 m) and was typically at or below the detection limit (~10 nM) from ~90 m to 700 m. At the time of sample collection, NO_3^- reduction rates increased with depth into the OMZ, peaking at ~9.5 nM N d^{-1} at 300 m[19], paralleling an increase in the abundance of sequences encoding Nar-type NO_3^- reductases in coupled metagenomes and metatranscriptomes (Fig. 6.1c). In contrast, aerobic NO_2^- oxidation peaked at 100 m (260 nM N d^{-1}) where trace O_2 was available and NO_2^- was abundant, before declining 20-fold with depth into the OMZ (Fig. 6.1c). However, NO_2^- oxidation rates are likely overestimated due to slight O_2 contamination in incubations[20]. These data highlight a transition to anoxia within the ETNP OMZ[15,19], with *in situ* $[O_2]$ at least an order of magnitude lower than the inhibitory threshold for NO_3^- reduction, denitrification, and anammox[25,26], consistent with micromolar accumulations of NO_2^- from NO_3^- reduction in this zone.

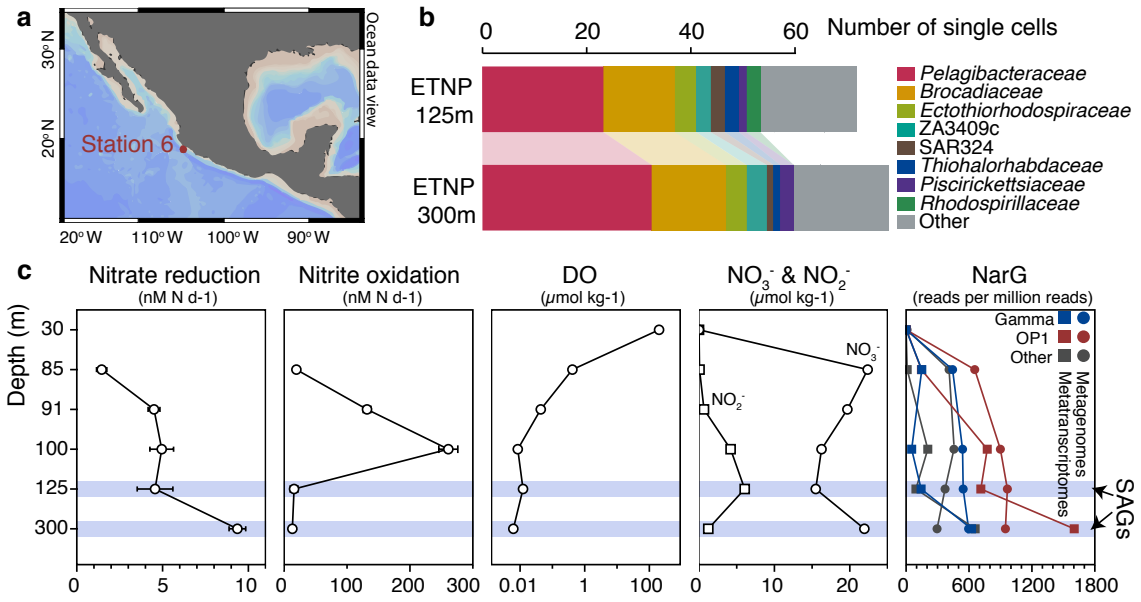


Figure 6.1: Site description and phylogenetic affiliation of single cells.

A, Location of station 6 (red) in the ETNP from which samples were obtained. B, Taxonomic classification of sorted single cells, based on their 16S rRNA genes. C, Nitrate reduction and nitrite oxidation rates relative to dissolved O_2 (DO), nitrate, and nitrite concentrations and *narG* read abundance in metagenomes and metatranscriptomes. Error bars represent standard error from triplicate measurements. Note that a $\log_{10}$ scale is used for the DO plot and that 0.01 μmol kg^{-1} represents the detection limit of the STOX sensor oxygen data presented here. DO at 300 m was below the detection limit.

6.3 METHODS

6.3.1 Single cell sample collection and sequencing

Samples for single cell sorting and single amplified genome (SAG) analysis were collected from the from the anoxic depths (Fig. 6.1) in station 6 at ETNP, on June, 2013. Additional ("control") oxygen rich samples were collected from a surface waters (1m) of the Gulf of Mexico (GoM) on May 29, 2012 aboard the *R/V Endeavor* (cruise EN509) at station 5. For cryopreservation of cells, triplicate 1 ml samples of bulk seawater (no

prefiltration) were gently mixed with 100 μl of a glycerol TE stock solution (20 ml 100X TE pH 8.0, 60 ml sterile water, 100 ml glycerol) and frozen at -80ºC. Samples for oxygen and nutrient measurements were collected on the same date and casts as those for single cell sorting described above and have been previously described[19].

Single amplified genomes (SAGs) were generated from individual bacterial cells[27], according to standard procedures in the Department of Energy Joint Genome Institute workflow[28]. Using PCR and Sanger sequencing of a ~470 bp region of the 16S rRNA gene, a total of 27 SAR11 classified SAGs were identified (Fig. 6.1B), for which libraries were prepared with the NexteraXT DNA Sample Prep kit (Illumina, San Diego, CA, USA) and sequenced on two runs of an Illumina MiSeq using a 500 cycle (paired end 250 x 250 bp) kit. Of the initial 27 SAGs, 8 were recovered in very low abundance in the read data or were removed due to potential contamination (>5%) as estimated with CheckM (see below) or the presence of 18S rRNA gene fragments, yielding the final set of 19 SAGs analyzed here (Table E1).

6.3.2 SAG sequence quality control assembly and functional gene annotation

Coupled reads were merged, when overlapping, using PEAR[29]. Both merged and un-merged reads were trimmed using SolexaQA++[30] with a PHRED score cutoff of 20 and a minimum fragment length of 50 bp. Illumina adaptors were clipped using Scythe (https://github.com/vsbuffalo/scythe) and reads were re-filtered for length (50 bp). Quality-trimmed reads were assembled with SPAdes[31]. Percentage of contamination and genome completeness were assessed based on recovery of lineage-specific marker gene sets using CheckM[32]. From the total of 27 SAG assemblies, 7 were excluded from the analysis due to low coverage (i.e., less than 70 kb) or the presence of 18S rRNA sequences and BLASTP top matches to eukaryotic sequences reflecting contamination. For the remaining SAGs that passed the original quality control thresholds (Table E1), when multiple fragments of a bacterial single-copy marker gene were identified, manual inspection of alignments revealed that multiplicity was due to assembly breaking points rather than contamination from divergent sequences, and such cases were retained for analysis. Evidence for contamination was detected in only one SAG, SAG A2 from the GoM, as multiplicity of divergent and nearly full-length marker genes. This SAG was

excluded from further analysis. For the final dataset of 19 SAGs, coding sequences were predicted on scaffolds longer than 500 bp with GeneMark.hmm[33] and 16S rRNA gene sequences were identified using RNAmmer[34]. 16S rRNA sequences identified in the assemblies (4/4 GoM SAGs, and 8/15 OMZ SAGs) were compared to the 470 bp 16S fragment obtained during the initial SAG screening and confirmed to be identical. As an additional quality control step, all predicted genes from the 19 SAGs were taxonomically annotated using MyTaxa[35] and the taxonomic distributions of adjacent genes in the concatenated assembly (10 gene windows) were inspected for possible contamination (Fig 6.2).

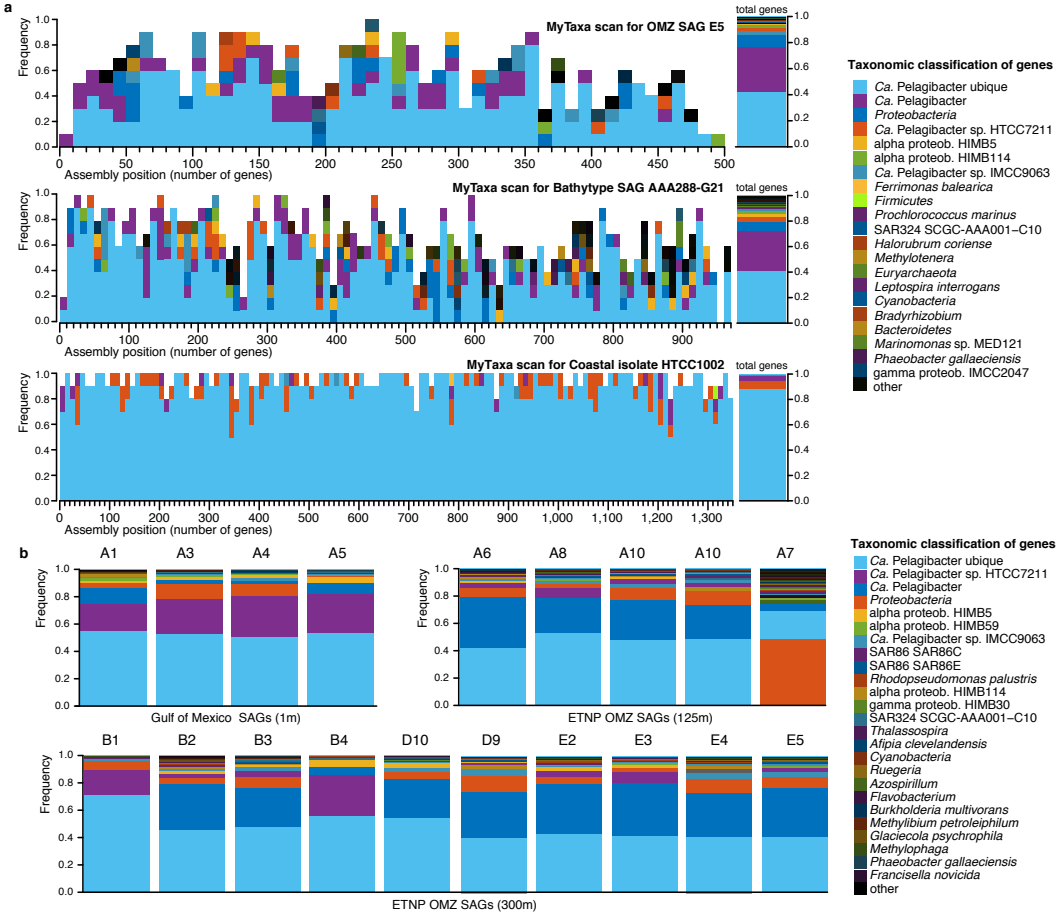


Figure 6.2: Evaluation of SAG contamination based on taxonomic affiliations.

a, Representative MyTaxa plots to test for contamination based on taxonomic affiliations of predicted genes. The MyTaxa algorithm predicts taxonomic affiliations and each gene

is assigned to the deepest taxonomic resolution (out of phylum, genus, and species) for which a high confidence value can be obtained (score 0.5). Each MyTaxa scan represents taxonomic distributions of all the predicted genes for one genome, given in windows of 10 genes, and sorted based on their position in the concatenated assembly of the genome (when a partial genome is used). White space in the histograms represents genes that could not be assigned to a given taxon due to (a) lack of BLASTP hits against the reference database (a collection of closed and draft genomes) or (b) lack of high confidence scores. Notice that for the representative OMZ SAG E5, more than 80% of the genes can be classified as *Candidatus* Pelagibacter (SAR11), with an additional 10% assigned to *Proteobacteria*. Note there are no genome representatives for this taxon (*i.e.*, SAR11 subclade IIa.A) in the database upon which MyTaxa is based. Similar results are obtained for the bathytype SAR11 SAG, as this genome also lacks representatives. The closed genome from a coastal isolate HTCC1002 is shown for comparison to demonstrate a typical pattern for cases when close relatives of the query genome are available in the reference database, as is the case for this isolate. **b**, Taxonomic classifications of genes from the 19 SAGs analyzed here. Each distribution was obtained from the MyTaxa scans performed for each SAG. The percentage of the total genes that could be taxonomically classified with MyTaxa was on average ~60%, and varied depending on the completeness of the genome (*i.e.*, partial genes are less likely to be assigned taxonomy with high confidence). Of the genes that could be classified, the majority (>90%) were classified to SAR11 taxa.

Predicted genes were functionally annotated using the blast2go pipeline[36] for assignment to metabolic pathways, and screened manually for evidence of anaerobic energy metabolism. Detected genes of anaerobic metabolism, including nitrate reductase (*nar*) genes, as well as terminal oxidase genes and the single-copy marker gene *rpoB,* were further verified using HMMER3 (http://hmmer.janelia.org/) with default settings and recommended cutoffs for a match against available Pfam models[37]. Statistics of SAG quality control, assemblies, and contamination testing are provided in Table E1.

6.3.3 Metagenome and metatranscriptome samples

ETNP OMZ metatranscriptomic and metagenomic datasets were generated via MiSeq Illumina sequencing in [19] and [38] respectively. Metagenome datasets from the ETSP were generated by Roche 454 pyrosequencing as previously described[18]. Metagenomes from the GoM samples were generated with the same protocols as the OMZ samples[18] and libraries were prepared and sequenced in two lanes on an Illumina HiSeq (150 bp paired reads). Metadata, sequencing statistics, and accession numbers of all analyzed metagenome and metatranscriptome datasets are provided in Table E3. All metagenomic and metatranscriptomic datasets were quality trimmed as described below for the SAG datasets. The metatranscriptomic datasets were further filtered to remove rRNA transcripts using the SortMeRNA algorithm[39]. 454 metagenomic datasets were filtered to remove duplicate sequences. The quality trimmed reads from the OMZ metagenomes (ETNP and ETSP), were assembled with IDBA[40] and genes were predicted on contigs longer than 500 bp with MetaGeneMark.hmm[33]. Taxonomic classification of metagenomic contigs was performed with MyTaxa[35]. *Nar* operons were identified on metagenomic contigs as described above for the SAG assemblies.

6.3.4 Phylogenetic placement of SAGs

The evolutionary relatedness of SAR11 SAGs was assessed using the identified full or almost full-length 16S rRNA gene sequences from the assembled SAGs. For the SAGs from which no full-length 16S rRNA fragments were assembled, the shorter fragments obtained during screening were used in pairwise comparisons with full-length sequence references. The 16S rRNA sequences from publicly available SAR11 genomes, as well as previously published 16S sequences[6,13] from subclades with no genome representatives, were included in the alignment to aid in the classification of the SAR11 subclades. Additionally genome representatives of divergent alphaproteobacteria classes, as well as a beta- and gammaproteobacterium were included to facilitate the rooting of the tree. Maximum likelihood phylogenetic reconstruction was performed with RAxML with 1000 bootstraps and the GTR model for nucleotides[41]. Additionally, Hidden Markov Models (HMMs) of 106 housekeeping genes found in single copy in bacterial genomes were used to identify marker genes in available SAGs and reference genomes

using HMMER3 (http://hmmer.janelia.org/) with default settings and the recommended cutoff[28]. The identified marker genes were aligned using Clustal Omega[43] and the protein alignments concatenated using Aln.cat.rb from the enve-omics collection (http://enve-omics.ce.gatech.edu/) to remove invariable sites and maintain protein coordinates. The concatenated alignment was used to build a maximum likelihood phylogeny with RAxML, using 1000 bootstraps, and the PROTGAMMAAUTO function, which identifies the best amino acid substitution model for each protein. SAGs where assigned to SAR11 subclades based on the consensus categorization of both 16S rRNA and marker gene phylogenies, in accordance with previously published subclade identification sequences[6,13]. OMZ-derived SAR11 SAGs from the SAR11 IIa lineage were further categorized as subclade IIa.A, to differentiate them from the currently available reference SAR11 IIa representative (HIMB058), classified here as subclade IIa.B. Average amino acid identities (AAI) were estimated as described previously[44].

6.3.5 Nar functional gene validation and phylogeny

Reference nitrate reductase and nitrite oxidoreductase protein sequences (n=697) representing divergent bacterial and archaeal phyla were downloaded from UniProt/Swiss-Prot[45], together with representatives of other DMSO family oxidoreductases (n=71), using as a guide the reference tree from [46]. From this 697-sequence set, 321 full-length NarG/NxrA sequences were selected to represent all the clades, along with the 71 additional non-NarG/NxrA proteins. The NarG/NxrA subset included the closest relatives to the SAG OP1 and Gamma-type Nar variants, as determined by BLAST. All protein sequences (n=392), including the full-length NarG identified in the SAGs, were aligned with Clustal Omega, and a maximum likelihood phylogeny was reconstructed with RAxML with 1000 bootstraps and the PROTGAMMAAUTO model. Partial fragments of the NarG protein were then added to the alignment using MAFFT's "addfragments"[47], and the Evolutionary Placement Algorithm (EPA) implemented in RAxML was used to place them within the reference tree[48]. The same procedure was followed for the phylogenetic reconstruction and placement of identified NarH protein sequences.

Quantification of *narG*-encoding reads from the metagenomes and metatranscriptomes was done using BLAST searches against a manually curated NarG database and the software ROCker[49]. Using Receiver-Operator Curve (ROC) analysis, ROCker identifies the most discriminant BLAST bit-score per position in a reference alignment (NarG database) given a certain read length by simulating *in silico* metagenomic datasets that include the reference genes. This strategy permits the accurate estimation of abundance of target genes in short-read datasets, minimizing false negatives and positives derived from closely related proteins or conserved domains, a critical challenge in the detection of *narG* due to the ubiquity of other closely related DMSO oxidoreductases. The NarG database was manually curated and confirmed by the phylogenetic reconstruction of all available nitrate reductase and nitrite oxidoreductase sequences and visual inspection of the multi-sequence alignment for conservation of known functional domains and motifs. The final NarG database consisted of 697 nitrate reductases/nitrite oxidoreductases (positive set) and 71 representative non-NarG/NxrA DMSO family proteins (negative set for identification of false positive BLAST matches). All datasets, as well as the ROCker models built for *narG* quantifications in metagenomes with different read lengths, are available at http://enve-omics.ce.gatech.edu/rocker/. Additionally, the model for the identification of *rpoB* fragments in metagenomes was used to estimate coverage of *rpoB* in metagenomes.

The abundance of *narG* sequences in meta-omic datasets was estimated as genomic equivalents for each sample, by normalizing the coverage of *narG* for the gene length (reads per nucleotide of *narG*), and dividing the normalized value by the *rpoB*-normalized coverage (reads per nucleotide of *rpoB*). In order to quantify the abundance of the *narG* variants (OP1, Gamma-type), protein fragments were predicted in all identified (from ROCker) *narG* reads using FragGeneScan[50] and placed in the reference DMSO tree using RAxML-EPA. The abundances of the OP1-type or Gamma-type variants were estimated based on the number of reads that were placed in the terminal or internal nodes of the aforementioned clades on the reference tree, using JPlace.to_iToL.rb from the enve-omics collection. BLASTP was used to map the reads against the reference NarG sequences, and the recruitment plots were constructed with the BlastTab.catsbj.pl and BlastTab.recplot.R scripts from the enve-omics collection.

Thus, the reported abundances of OP1 and Gamma-type *narG* in metagenomes/metatranscriptomes are based on phylogenetic assignment of *nar* reads, rather than a strict sequence similarity cutoff. In order to estimate a lower limit for the abundance of NarG sequences presumably encoded by SAR11 genomes, the number of reads with more than 95% nucleotide identity to the reference NarG sequences found in the SAGs was estimated.

6.3.6 Quantification of SAR11 clades in metagenomes and metatranscriptomes

For each metagenome/metatranscriptome, reads potentially derived from SAR11 genomes were identified by a competitive BLAST best-match approach. A custom database was built using all available closed genomes from NCBI-ftp (2638 bacterial, 165 archaeal) and 39 genome representatives of the SAR11 lineage, including 20 published isolate or SAG sequences and the 19 SAG sequences produced in this study (Table E1 and E2). Metagenomic and metatranscriptomic reads (predicted orfs with FragGeneScan) were then compared against the database using BLASTP, and the subset of reads with a best match against any of the SAR11 genomes and an e-value < 0.001 was classified as "SAR11 reads". To quantify the relative abundance of distinct SAR11 subclades, the "SAR11 reads" were further classified as follows. We used the coverage of universal marker genes that could be found in all the subclades to more accurately estimate the abundance of distinct subclades and overcome both the biased representation of SAR11 subclades in the available genomes, and the partial nature of SAG genomes. For all 39 available SAR11 genomes, 5707 orthologous genes (OGs) were identified by reciprocal best match and Markov Clustering with inflation 1.5 using ogs.mcl.rb from the enve-omics collection. From the identified OGs, 507 were represented at least once in each of the 8 subclades. All metagenomic and metatranscriptomic reads (SAR11 subsets) were mapped against the database containing all protein sequences from the 507 OGs (which were tagged according to subclade of origin) using the BLASTX option from Diamond[51] and only the best matches for each read were kept. The coverage of each OG for each subclade was estimated based on that competitive best match result, normalized for the gene length (reads per bp of each OG), and the average coverage of all 507 OGs was used to estimate the abundance of subclades. Additionally, the number of *rpoB* reads for each metagenome

was identified (for either the total dataset or the subset of the SAR11 reads), and the coverage of *rpoB* was used as a normalization factor to estimate the abundance of SAR11 subclades over the total bacterial community.

6.3.7 Functional characterization of SAR11 nar operons

A previously constructed NO_3^- reductase deficient *Escherichia coli* strain[52] was used as the genetic system for heterologous expression of SAR11 *nar* genes. We used whole genome sequencing (Illumina MiSeq) to confirm that this strain lacked all three NO_3^- reductases *(*Δ*narGI* Δ*napAB narZ*::Ω). The phenotype of this strain, hereafter referred to as the triple mutant, was verified by a lack of NO_2^- production and an absence of growth with NO_3^- under anaerobic conditions, compared to the wild type MC4100 *E. coli* strain.

Complete sequences from one OP1-type, and one Gamma type *nar* operon, containing upstream and downstream sequences, were identified from the ETNP-300 m and ETNP-120 m metagenomes (see above). These sequences were confirmed to be identical to the operons in SAG A7 (which was lacking part of the N-terminus of the NarG gene). Purified DNA from the ETNP-300 m and ETNP-120 m metagenomic samples was used as template for PCR amplification. The resulting PCR products were gel purified, assembled and cloned into pBbA1K, a low copy vector including the IPTG-inducible pTrc promoter [53] by In-Fusion cloning (Clontech, Mountain View, CA). The cloning reactions were transformed into TOP10 cells, and inserts were sequence-verified by Pacbio sequencing (Pacific Biosciences, Menlo Park, CA). The final *nar* sequences were identical (OP1 operon, and NarG,I proteins of Gamma operon) or nearly identical with silent substitutions (99% and 98% aa identity for the Gamma-type NarG and H proteins) compared to the sequences from SAG A7. Correct clones were isolated for each operon type, and purified plasmid was used to electroporate the triple mutant *E. coli* strains described above to generate recombinant strains expressing the heterologous *nar* operons for functional characterization.

8 replicate clones (per recombinant strain) were grown aerobically on 96-well plates in 70 μl Luria-Bertani (LB) broth supplemented with 30 mM NO_3^- and various IPTG concentrations, and tested for NO_2^- production via the Griess reaction. Clones that

produced nitrite were further tested for anaerobic growth dependent on NO_3^-. The clones were first induced in LB medium with 0.5mM IPTG for 5h, and 20 μl of inoculum was subsequently introduced in gas tight tubes under N_2 atmosphere. The medium was prepared as previously described[54], composed from potassium phosphate buffer (100 mM, pH 7.4), 15 mM $(NH_4)_2SO_4$, 9 mM NaCl, 2 mM $MgSO_4$, 5 μM Na_2MoO_4, 10 μM Mohr's salt, 100 μM $CaCl_2$, 0.5% casaminoacids and 0.01% thiamine. Glycerol (40 mM) was used as the sole carbon, and NO_3^- was added at 30 mM. IPTG (0.5 mM), kanamycin (30μg/ml) and streptomycin (30 μg/ml) were used with the recombinant strains. Samples for NO_3^- and NO_2^- concentrations were obtained at regular time intervals during incubations, filtered through 0.2 μm porosity filters and injected into a Dionex DX ion chromatography unit with the Dionex IonPac AS14A analytical column[55]. Growth in incubations was assessed as optical density (OD; 600 nm). Growth curve data from replicated cultures (triplicate) were fitted to a logistic model with variables r (specific growth rate), P_0 (initial population), and K (carrying capacity), using nonlinear least-squares estimates and prediction of OD per time point with confidence intervals.

6.4 RESULTS AND DISCUSSION

6.4.1 Diverse SAR11 single cell genomes from anoxic waters

Samples for SAG analysis were obtained from two depths in the anoxic zone: at 125 m at the NO_2^- maximum (6 μM), and at 300 m in the core of the NO_3^- reduction zone. Single prokaryotic cells were isolated by fluorescence-activated sorting, subjected to genome amplification[56], and screened by 16S rRNA gene fragment (470 bp) PCR and Sanger sequencing. From this screen, 23% and 32% of SAGs from 125 and 300 m, respectively, were confidently assigned to the SAR11 Family *Pelagibacteraceae* (Fig. 6.1b), thus confirming SAR11's substantial numerical abundance in the OMZ. From this SAR11 subset, 10 SAGs from 125 m and 12 SAGs from 300 m were randomly selected for shotgun sequencing (Illumina), along with 5 technical control SAR11 SAGs from the oxic surface waters of the Gulf of Mexico (GoM). Following sequencing, quality filtering, and assembly, a total of 19 SAGs were used for analysis: 15 OMZ SAGs (5 from 125 m,

10 from 300 m) and 4 GoM control SAGs (Table E1). These genomes exhibited varying levels of completeness (~2-90%; average 30%) and no detectable contamination (Fig. 6.2), as assessed by the presence of single-copy housekeeping genes[32,42], 16S rRNA gene identities, and the taxonomic assignment of SAG contigs.

The identified SAGs represented a diverse and novel SAR11 community in the OMZ. Phylogenetic reconstructions based on either 16S rRNA genes or single-copy housekeeping proteins placed the 19 SAGs in 5 subclades of SAR11 (Fig. 6.3a). Average amino acid identity (AAI) comparisons among all available SAR11 genomes (Table E2) further corroborated this classification, placing (i) 7 OMZ SAGs within the previously uncharacterized deep-branching monophyletic group of subclade IIa (hereafter designated subclade IIa.A), distinct (>5% 16S divergence) from SAG HIMB058 from the tropical North Pacific (hereafter designated subclade IIa.B), (ii) 3 OMZ SAGs within the deep-branching subclade IIb, (iii) 2 OMZ SAGs within subclade Ic, which includes recently described single-cell genomes from the bathypelagic ocean[6], (iv) 2 OMZ and all 4 GoM surface SAGs within subclade Ib, which thus far lacks genome representatives, and (v) OMZ SAG A7 as most closely related to HIMB59, a member of the divergent SAR11 subclade V[8,57,58]. Note that the exact placement of subclade V in the SAR11 phylogeny is unstable depending on the marker gene and outgroup used[59,60]. The average estimated genome size of OMZ SAGs was 1.33 Mbp, consistent with prior reports of genome streamlining in SAR11.

6.4.2 OMZ SAR11 abundance peaks under oxygen depletion

To estimate the *in situ* abundance and activity of OMZ SAR11, metagenome and metatranscriptome reads from OMZ sites and from diverse oxic ocean regions (Table E3) were recruited to 39 available SAR11 genomes (Table E1 and E2). Metagenomic read recruitment, performed essentially as described previously[61], showed that each OMZ SAR11 subclade represents a sequence-discrete (and hence tractable) population, but with each population encompassing substantial intra-population variation (~92-100% average nucleotide identity between members of the population vs. <90% between populations), as well as gene content variability. We therefore estimated SAR11 abundance at the subclade level, based on the average coverage of 507 genes shared

between genomes from all SAR11 subclades. Based on this analysis, SAR11 subclades Ic, IIa.A, and IIb together comprised about 10 to 30% of the bacterial community in ETNP and ETSP metagenomes and metatranscriptomes from depths with undetectable O_2 (Fig. 6.3b, 2c), consistent with the high abundance of SAR11 in the pool of cells sorted for SAG analysis (Fig. 6.1b). Subclade IIa.A, composed exclusively of 7 SAGs from this study, was particularly abundant, making up to 15% of the community in anoxic samples. All OMZ subclades were absent from or much less abundant (<5%) in metagenomes from oxic sites, including those from above the ETNP OMZ (Fig. 6.3b). Together, these results identify newly described SAR11 subclades whose distribution is linked to an oxygen-depleted niche.

6.4.3 Metabolic adaptations to low oxygen in SAR11 genomes

OMZ and GoM SAGs were then analyzed for evidence of microaerobic or anaerobic metabolism. Surprisingly, in 8 of the 15 OMZ SAGs, belonging to SAR11 subclades Ic, IIa.A, IIb and V, protein family-based classification detected genes encoding the respiratory NO_3^-reductase (Nar) of the DMSO reductase superfamily (Fig. 6.3a). Evidence of a complete canonical *nar* operon (*narGHJI*) –encoding the α subunit that catalyzes NO_3^- reduction to NO_2^- (NarG), the iron-sulfur-containing β subunit (NarH) that transfers electrons to the molybdenum cofactor of NarG, the transmembrane cytochrome b-like γ subunit (NarI) involved in electron transfer from membrane quinols to NarH, and the NarJ chaperone involved in enzyme formation– was found within a single assembled contig in 4 SAGs (A6, E4, D9, A7), while partial *narG* and *narH* fragments were identified in another 4 SAGs (Fig. 6.4). In all SAR11 SAGs containing *nar* on a contig, we identified other genes upstream or downstream on the same contig taxonomically assigned to SAR11 reference genomes, further confirming the association of *nar* with SAR11. Genes encoding the NO_3^-/ NO_2^- transporter NarK and proteins for biosynthesis of the essential molybdenum cofactor (*moeA, mobA*) were also identified in 8 and 5 of the SAGs, respectively (Table E1). In only 4 of the 15 OMZ SAGs were Nar or cofactor synthesis genes not detected, presumably due to sequencing gaps (completeness of these SAGs: 4-20%). In contrast, these genes were not detected in any of the 4 control SAGs from the oxic GoM, despite high completeness of those genomes (average 61%).

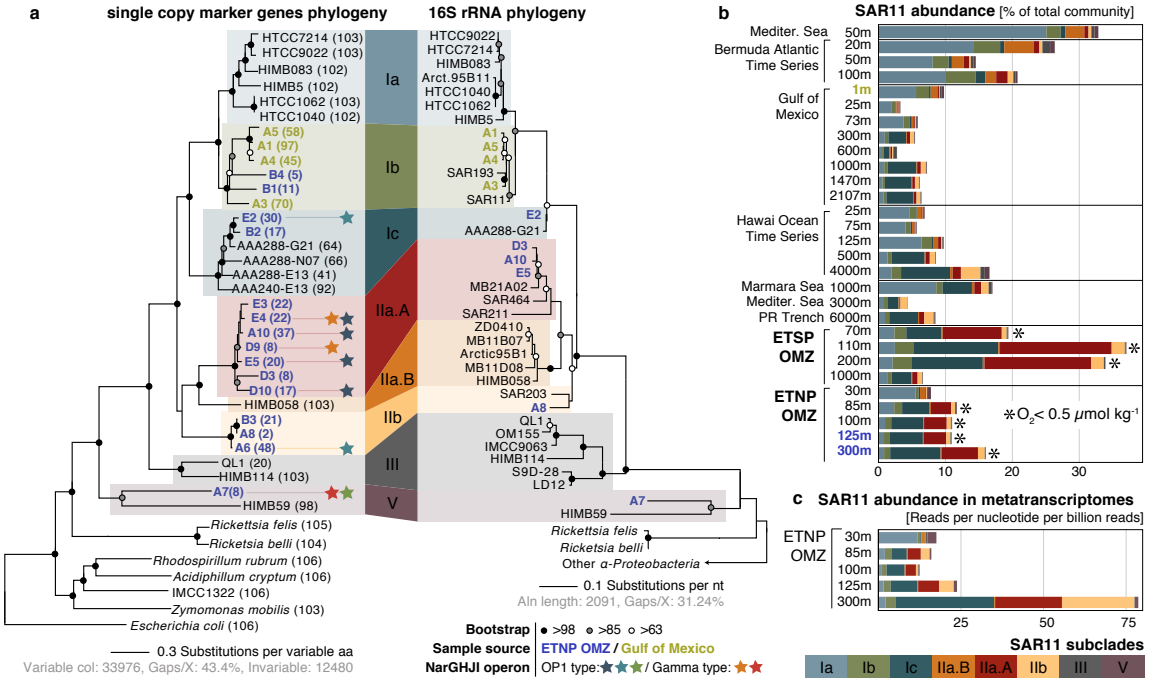


Figure 6.3: Diversity, abundance, and transcription of nitrate-reducing SAR11.

A, Maximum likelihood phylogeny based on the concatenated alignment of single copy housekeeping (left) and 16S rRNA (right) genes in SAGs from this study, SAR11 and representative alphaproteobacterial genomes. Values in parentheses denote the number of housekeeping genes used per genome. For the 16S-based tree, only full-length sequences from the genomes in the left tree were included. Star symbols of the same color represent closely related *narG* genes (>97% aa identity), encoding the catalytic subunit of the respiratory nitrate reductase of the DMSO family. B, Abundance of SAR11 subclades (left) in selected oceanic metagenomes. Note that the major *nar*-encoding clade IIa.A peaks in abundance at oxygen-depleted OMZ depths. C, Normalized average coverage of SAR11 subclades in ETNP metatranscriptomes. Transcription by *nar*-encoding lineages increases from the base of the oxycline (85 m) to spike at the OMZ core (300 m), but is negligible in the overlying oxic zone (30 m).

Genes encoding for downstream steps of denitrification or other dissimilatory anaerobic metabolisms were not found in any of the SAGs. However, in contrast to all previously analyzed SAR11 genomes, 3 of the OMZ SAGs, all from subclade IIa.A, also contained genes encoding high-affinity O_2-utilizing *bd*-type terminal oxidases (Table E1). Compared to the coxI-type oxidases present in all known SAR11 genomes, including the OMZ SAGs analyzed here, *bd*-type oxidases have a much higher affinity for O_2 (3-8 nM), suggesting a potential for microaerobic respiration by OMZ SAR11. These results provide the first indication of adaptation to low oxygen in SAR11 and the ability to respire NO_3^- to NO_2^- in the absence of oxygen, consistent with the distribution of these bacteria in the OMZ water column.

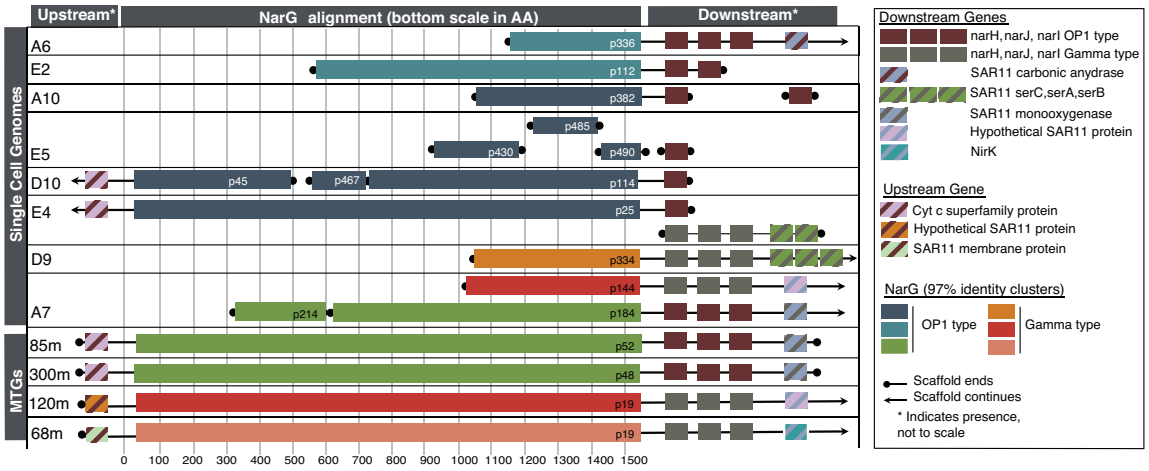


Figure 6.4: *nar* genes encoded by SAR11 populations of OMZs.

nar operon and adjacent genes identified in SAR11 single amplified genomes (SAGs) from the ETNP OMZ, and in assemblies from the 85 m and 300 m ETNP OMZ metagenomes. *narG* sequences with at least 97% amino acid similarity are represented with the same color.

6.4.4 Multiple divergent nitrate reductases in OMZ SAR11

Phylogenetic placement of all identified *narG* and *narH* genes and partial fragments revealed two divergent *nar* variants in OMZ SAGs (Fig. 6.5): (i) an "OP1-type" in which all 4 *nar* genes and an upstream cytochrome c protein were most similar (56-78% amino acid identity) to homologs from '*Candidatus* Acetothermus autotrophicum', a putative anaerobic acetogen of the candidate bacterial phylum OP1[62], and (ii) a "Gamma-type" variant most similar (51-78% identity) to Nar from a denitrifying *Gammaproteobacteria* endosymbiont (*Ca.* Vesicomyosocius okutanii strain HA)[63]. At least two of the OMZ SAR11 SAGs from subclade IIa.A, as well as SAG A7 from subclade V, encoded both OP1- and Gamma-type *nar* variants, suggesting that divergent *nar* copies (~42% AAI) co-occur in the same genome. Multiple *nar* operons per genome have been reported for diverse bacteria and are hypothesized to be related to adaptation to different oxygen conditions, with one variant constitutively expressed at low baseline levels[64–66]. For both OP1- and Gamma-type variants, the sequence divergence among recovered sequences was consistent with the phylogenetic placement of the SAGs. For example, OP1-type *narG* fragments represented 3 distinct 97% amino acid identity clusters (Fig. 6.3a). Sequences from clade IIa.A SAGs fell within the same cluster, sharing ~96.5% identity with sequences of the closely related Ic and IIb subclades, and ~90% with sequences from the more distant A7 SAG (Fig. 6.5). This pattern suggests diversification of *nar* operons in parallel with its genomic background, and also confirms that these sequences are not a systemic contaminant.

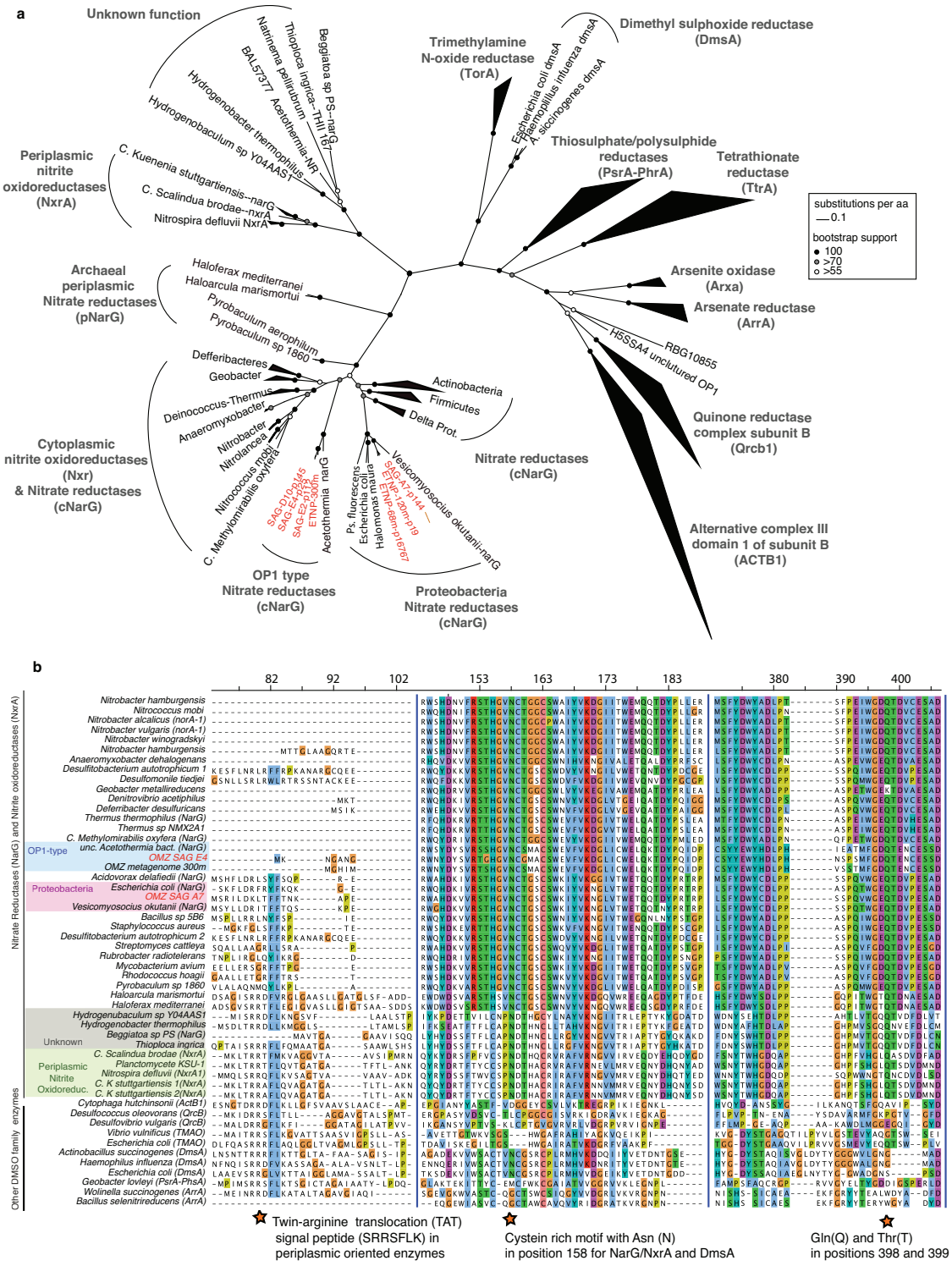


Figure 6.5: SAR11 NarG belong to the DMSO superfamily of oxidoreductases.

A, Phylogenetic reconstruction of NarG sequences identified in the SAR11 SAGs and metagenomic SAR11 contigs (ETNP prefix), along with reference NO_3^- reductase and

NO_2^- oxidoreductase enzymes. The tree also includes representative diverse DMSO oxidoreductases for direct comparison with the NarG/NxrA enzymes. **b**, Alignment of NarG sequences from OMZ SAR11 with representative sequences from the DMSO superfamily of oxidoreductases. The protein motifs in the second and third panels are present in all functional nitrate reductases (NarG) and nitrite oxidoreductases (NxrA) but not in closely related enzymes of the DMSO superfamily. The first panel shows the presence/absence of the TAT signal peptide (SRRSFLK), whose presence typically denotes a protein excreted to the outer membrane. SAR11 NarG is instead oriented toward the cytoplasm (lack of TAT). The second panel shows the cysteine-rich motif typically found in the N-terminus of the type-II DMSO superfamily oxidoreductases [67] and believed to enable the formation of a [4Fe–4S] cluster in these proteins. The Asn (N) in position 158 of the alignment is typically found in catalytic subunits of nitrite reductases and DMSO oxidoreductases (DmsA) but not in other DMSO family enzymes. The third panel shows the Gln(Q) and Thr(T) in positions 398 and 399 within the putative substrate entry channel of the protein, which differentiate the Nar proteins from all other oxidoreductases of the DMSO family.

6.4.5 Sequence-based and experimental characterization of SAR11 nitrate reductases

We sought to further characterize the biochemical function of SAR11 *nar* genes. Phylogenetic reconstruction based on 392 proteins of the diverse DMSO superfamily revealed that both OP1- and Gamma-type NarG fall within the clade of membrane-bound cytoplasm-oriented NO_3^- reductases (Nar) and NO_2^- oxidoreductases (Nxr), and were most closely related to Nar from known NO_3^- reducing bacteria (Fig. 6.5)[68]. The lack of a TAT peptide motif at the N-terminus corroborated the probable cytoplasmic orientation of the NarG active site[69], similar to experimentally verified Nar in *Escherichia coli*[70]. Additionally, the identified NarG sequences contain diagnostic functional domains found in NarG but not in other oxidoreductases of the DMSO reductase superfamily (Fig. 6.5)[68].

In order to verify NO_3^- reduction potential in SAR11 we introduced full-length SAR11 *nar* operons into a NO_3^- reductase-deficient *Escherichia coli* mutant and tested for enzyme activity. The Gamma-type *nar* operon was successfully expressed in *E. coli,*

yielding Nar proteins of the predicted size range and enabling growth of the mutant under anoxic conditions in the presence of NO_3^-, coupled with simultaneous NO_3^- reduction to NO_2^- (Fig. 6.6), thereby providing direct evidence for the function of this enzyme *in vivo*. The OP1-type operon did not reverse the *E. coli* mutant phenotype, presumably due to the much greater divergence of this variant from the *E. coli nar* operon. Given the high similarity of Nar and Nxr protein sequences[71–73], and the reversibility of the NO_3^- reduction reaction, it is possible that either or both OP1- and Gamma-type proteins could also function *in situ* to aerobically oxidize NO_2^-. While enticing, this possibility is remote given the experimental and phylogenetic evidence, a positive relationship between NO_3^- reduction rates and the abundance of OP1 and Gamma-type genes and transcripts in the anoxic OMZ depths (Fig. 6.1c), and prior results showing O_2 sensitivity of OP1-type *nar* transcription[25]. Rather, the results strongly suggest that the identified SAR11 *narG* genes encode functional NO_3^- reductases.

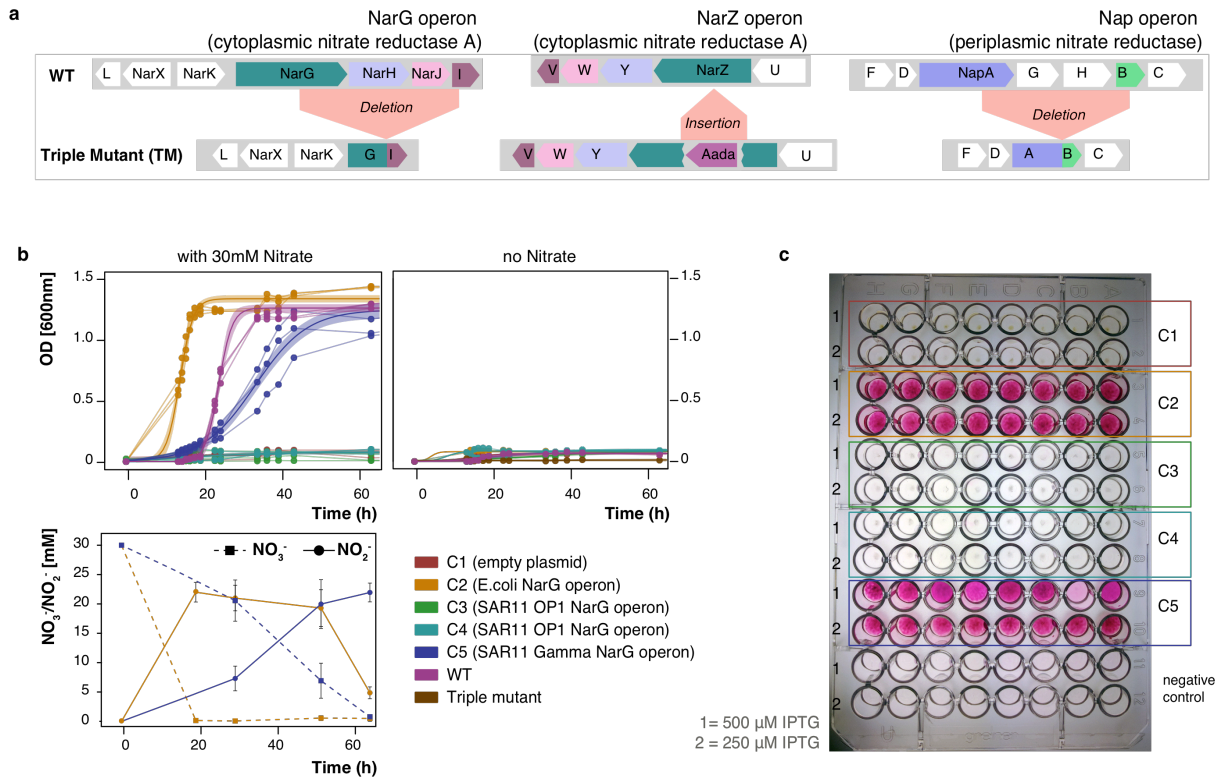


Figure 6.6: Functional characterization of the SAR11 *nar* operons.

a, Genotype of the *E. coli* triple mutant confirmed by whole genome sequencing. The triple mutant lacks complete functional operons of all three NO_3^- reductases, and thus is incapable of NO_3^- reduction. **b**, Anaerobic growth of triple mutant clones, complemented with the SAR11 *nar* operons. For each strain three independent clones were monitored, and data from the replicate growth curves were fitted into a logistic model. Shaded areas represent the 95% confidence intervals of optical density readings (OD600nm) in the fitted logistic growth models. NO_3^- and NO_2^- were measured in parallel with ion chromatography. Note that the Gamma-type SAR11 operon complements the triple mutant phenotype, growing anaerobically by reducing NO_3^- to NO_2^-. *E. coli* encodes functional nitrite reductases, thus the accumulated NO_2^- can be further reduced to ammonia, accounting for the non-stoichiometric NO_2^- production. **c**, Whole cell NO_2^- production assays under aerobic conditions. Eight independent clones (columns A-H) of each type (C1-C5) were inoculated in LB supplemented with 30mM NO_3^- and different IPTG concentrations, and the well plate was incubated for 2 days at room temperature. Griess reagent was added, and development of pink color indicated NO_2^- production.

6.4.6 SAR11 *nar* genes and transcripts are abundant in anoxic OMZ waters

We next examined the abundance of SAR11-affiliated *nar* genes within the OMZ to evaluate the contribution of SAR11 cells to NO_3^- reduction. We first identified *nar* sequence reads in OMZ metagenomes using a similarity search-trained model that discriminates NO_3^- reductase (or NO_2^- oxidoreductase) reads from those of other genes of the DMSO superfamily. These *narG* reads were then classified within a reference phylogeny containing 320 NarG proteins, including OP1- and Gamma-type sequences. Remarkably, the majority of *narG* reads from OMZ metagenomes were classified as OP1- or Gamma-type NO_3^- reductases (Fig. 6.7, Fig. 6.8), with the two variants accounting for 70% of total *narG* sequences at anoxic depths. The average number of *nar* genes per cell (i.e., genome equivalents) was estimated by comparing the abundance of *nar* sequences with those of *rpoB*, a universal single-copy gene. Based on those estimations, Gamma and OP1 *nar* variants occur in up to 61 and 85% of OMZ bacteria, respectively (Fig. 6.8), assuming each *nar* type occurs once per genome. Such high values are striking but consistent with prior results based on BLAST-based taxonomic assignments[18]. These values also exceed the estimated SAR11 abundances in the metagenomes, or those calculated directly from SAG 16S screening (up to 32% of the community), indicating that these gene variants occur in multiple copies per genome or in diverse bacteria. Metagenomic evidence suggests that the majority of these *nar* operons are found in SAR11 genomes within the OMZ. First, while our SAG collection captured only a fraction of total *nar* diversity, additional *nar* operons were identified in metagenomic contigs classified as SAR11 (Fig. 6.4). Second, the majority of the metagenomic *narG* reads showed >95% nucleotide identity with the *narG* genes encoded by the SAGs, suggesting that SAR11 cells are among the major contributors of NO_3^- reductases in the OMZ (Fig. 6.8).

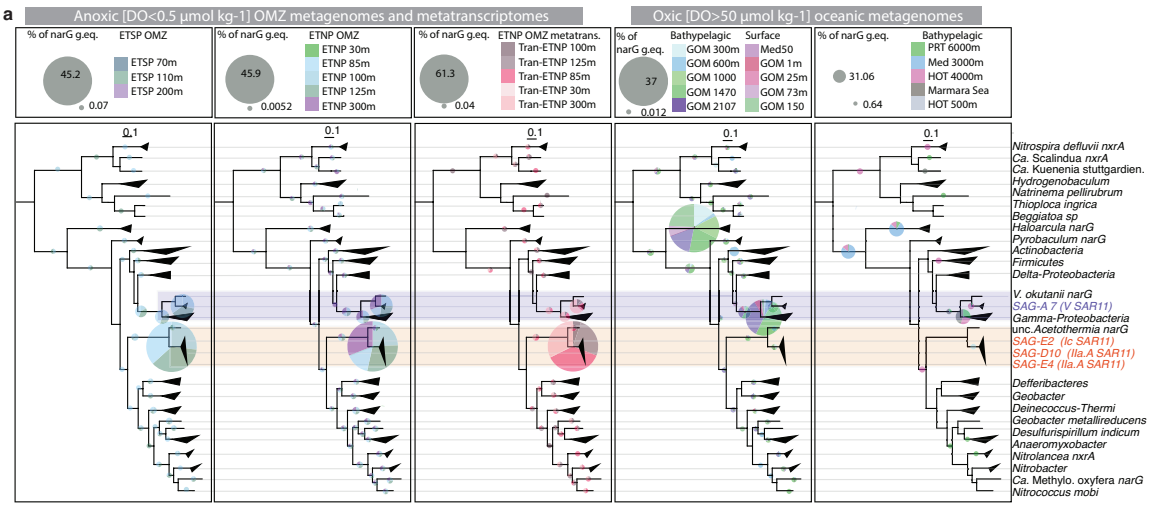


Figure 6.7: Relative abundance of *narG* variants in OMZ various ocean datasets. Relative abundance and diversity of NarG/NxrA enzymes as revealed by phylogenetic placement of identified *narG* metagenomic reads (colored pies). All identified short metagenomic *narG* reads from various oceanic metagenomes were placed within a reconstructed reference NarG tree in order to estimate the abundance of the different *narG* variants. The results of the placement are presented in 5 separate trees, based on the origin of the analyzed metagenomic reads for clarity. In each of the 5 trees, the colored pies represent the abundance (normalized for dataset size) of the short metagenomic reads clustering in the respective node. Specifically, the pie radius reflects read abundance as a percentage of the total *narG* genome equivalents identified, with the size of grey pies in the legends representing the highest and lowest relative abundance, respectively. The reference tree is the same as in Fig. 6.5A. Scale bars represent substitutions per amino acid. Notice that the two *narG* variants affiliated with the SAR11 SAGs (highlighted in orange for the OP1 type and blue for the Gamma type) are only abundant in the metagenomes and metatranscriptomes from the OMZ, where they comprise more than 70% of the total *narG* read pool, as can also be observed in Fig. 6.8.

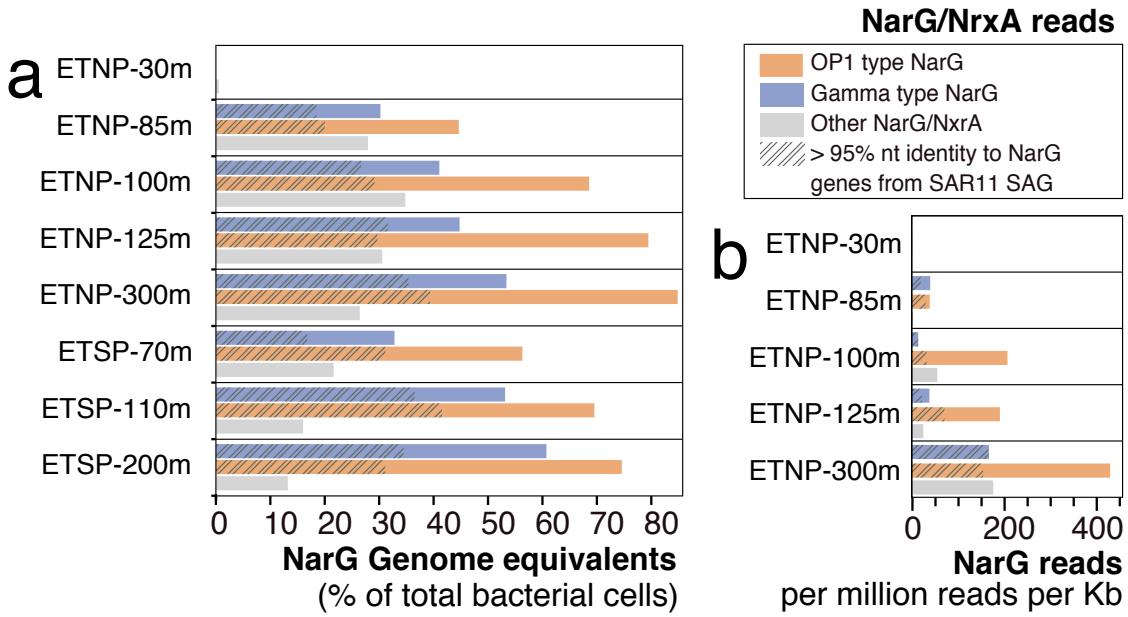


Figure 6.8: Diversity, abundance, and transcription of *nar* in the OMZ.

A, Relative abundance of NarG/NxrA enzymes in OMZ metagenomic datasets. Abundance was normalized to the *rpob* gene abundance and thus represents genome equivalents, or the portion of OMZ bacterial cells that encode the enzyme. B, Relative expression of NarG/NxrA proteins in the ETNP transcriptomes.

Metatranscriptome sequencing confirmed that SAR11-affiliated *nar* genes are transcribed in the OMZ. The abundance of both OP1- and Gamma-type variants in ETNP metatranscriptomes increased steadily from the lower oxycline (85 m) to the OMZ core (300 m), directly paralleling the abundance of the respective genes and the depth trend in NO_3^- reduction rates (Fig. 6.8). Notably, within the ETNP OMZ, an average of 39% of all *narG* transcripts shared >95% nucleotide identity with the OP1- or Gamma-type sequences detected in SAR11 SAGs (Fig. 6.8), a conservative lower-bound estimate of the contribution of SAR11 bacteria to the total *nar* transcripts within the OMZ. Accordingly, within the anoxic OMZ depths, *nar* genes are among the most transcriptionally active genes in the SAG genomes. The high transcriptional activity of SAR11 *nar* operons, interpreted alongside their distribution relative to NO_3^- reduction rates, suggests that SAR11 bacteria contribute substantially to community NO_3-respiration.

6.5 CONCLUSIONS

Collectively, our findings identify diverse and abundant SAR11 lineages whose genome content and environmental distribution reflect adaptation to an anoxic niche, unlike all other SAR11 bacteria characterized to date. The experimentally verified NO_3^- reductase activity in the Gamma-type SAR11 *nar* variant, along with the high expression levels of divergent SAR11 *nar* genes in the functionally anoxic core of the OMZ, suggest that persistence in this niche is linked to NO_3^- respiration, consistent with the fundamental importance of this process in OMZs. Nitrate respiration in OMZs constitutes the primary mode for organic carbon mineralization and the main production route of NO_2^-, a critical substrate for the major nitrogen loss processes of anammox and denitrification. The presence and activity of *nar* operons in SAR11, as well as the high abundance of *nar*-associated SAR11 clades in the OMZ, implicate these versatile organisms as major contributors to the initiation of OMZ nitrogen loss. Together, these findings redefine the ecological niche of one of the planet's most dominant group of organisms, providing a set of genomic references to establish SAR11 as a model for studies of nitrogen and carbon cycling in OMZs.

6.6 REFERENCES

1. Brown, M. V., Schwalbach, M. S., Hewson, I. & Fuhrman, J. A. Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series. *Environ. Microbiol.* **7,** 1466–1479 (2005).
2. Carlson, C. A. *et al.* Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. *ISME J.* **3,** 283–295 (2008).
3. Eiler, A., Hayakawa, D. H., Church, M. J., Karl, D. M. & Rappé, M. S. Dynamics of the SAR11 bacterioplankton lineage in relation to environmental conditions in the oligotrophic North Pacific subtropical gyre. *Environ. Microbiol.* **11,** 2291–2300 (2009).
4. Morris, R. M. *et al.* SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420,** 806–810 (2002).
5. Salter, I. *et al.* Seasonal dynamics of active SAR11 ecotypes in the oligotrophic Northwest Mediterranean Sea. *ISME J.* **9,** 347–360 (2015).
6. Thrash, J. C. *et al.* Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. *ISME J.* **8,** 1440–1451 (2014).
7. Giovannoni, S. J. *et al.* Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309,** 1242–1245 (2005).
8. Grote, J. *et al.* Streamlining and Core Genome Conservation among Highly Divergent Members of the SAR11 Clade. *mBio* **3,** e00252–12 (2012).
9. Tripp, H. J. The unique metabolism of SAR11 aquatic bacteria. *J. Microbiol. Seoul Korea* **51,** 147–153 (2013).
10. Konstantinidis, K. T., Braff, J., Karl, D. M. & DeLong, E. F. Comparative Metagenomic Analysis of a Microbial Community Residing at a Depth of 4,000 Meters at Station ALOHA in the North Pacific Subtropical Gyre. *Appl. Environ. Microbiol.* **75,** 5345–5355 (2009).
11. Swan, B. K. *et al.* Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333,** 1296–1300 (2011).
12. King, G. M., Smith, C. B., Tolar, B. & Hollibaugh, J. T. Analysis of composition and structure of coastal to mesopelagic bacterioplankton communities in the northern gulf of Mexico. *Front. Microbiol.* **3,** 438 (2012).
13. Vergin, K. L. *et al.* High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. *ISME J.* **7,** 1322–1332 (2013).
14. Paulmier, A. & Ruiz-Pino, D. Oxygen minimum zones (OMZs) in the modern ocean. *Prog. Oceanogr.* **80,** 113–218 (2009).
15. Tiano, L., Garcia-Robledo, E. & Revsbech, N. P. A New Highly Sensitive Method to Assess Respiration Rates and Kinetics of Natural Planktonic Communities by Use of the Switchable Trace Oxygen Sensor and Reduced Oxygen Concentrations. *PLoS ONE* **9,** e105399 (2014).
16. Kalvelage, T. *et al.* Nitrogen cycling driven by organic matter export in the South Pacific oxygen minimum zone. *Nat. Geosci.* **6,** 228–234 (2013).
17. Stewart, F. J., Ulloa, O. & DeLong, E. F. Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ. Microbiol.* **14,** 23–40 (2012).

18. Ganesh, S., Parris, D. J., DeLong, E. F. & Stewart, F. J. Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. *ISME J.* **8,** 187–211 (2014).
19. Ganesh, S. *et al.* Size-fraction partitioning of community gene transcription and nitrogen metabolism in a marine oxygen minimum zone. *ISME J.* **9,** 2682–2696 (2015).
20. Ulloa, O., Canfield, D. E., DeLong, E. F., Letelier, R. M. & Stewart, F. J. Microbial oceanography of anoxic oxygen minimum zones. *Proc. Natl. Acad. Sci.* **109,** 15996–16003 (2012).
21. Codispoti, L. A. *et al.* The oceanic fixed nitrogen and nitrous oxide budgets: Moving targets as we enter the anthropocene? *Sci. Mar.* **65,** 85–105 (2001).
22. Gruber, N. in *The Ocean Carbon Cycle and Climate* (eds. Follows, M. & Oguz, T.) 97–148 (Springer Netherlands, 2004).
23. Stewart, F. J., Sharma, A. K., Bryant, J. A., Eppley, J. M. & DeLong, E. F. Community transcriptomics reveals universal patterns of protein sequence conservation in natural microbial communities. *Genome Biol.* **12,** R26 (2011).
24. Lüke, C., Speth, D. R., Kox, M. A. R., Villanueva, L. & Jetten, M. S. M. Metagenomic analysis of nitrogen and methane cycling in the Arabian Sea oxygen minimum zone. *PeerJ* **4,** e1924 (2016).
25. Dalsgaard, T. *et al.* Oxygen at Nanomolar Levels Reversibly Suppresses Process Rates and Gene Expression in Anammox and Denitrification in the Oxygen Minimum Zone off Northern Chile. *mBio* **5,** e01966–14 (2014).
26. Kalvelage, T. *et al.* Oxygen Sensitivity of Anammox and Coupled N-Cycle Processes in Oxygen Minimum Zones. *PLoS ONE* **6,** e29299 (2011).
27. Raghunathan, A. *et al.* Genomic DNA amplification from a single bacterium. *Appl. Environ. Microbiol.* **71,** 3342–3347 (2005).
28. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499,** 431–437 (2013).
29. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinforma. Oxf. Engl.* **30,** 614–620 (2014).
30. Cox, M. P., Peterson, D. A. & Biggs, P. J. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* **11,** 485 (2010).
31. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).
32. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25,** 1043–1055 (2015).
33. Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* **38,** e132 (2010).
34. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35,** 3100–3108 (2007).
35. Luo, C., Rodriguez-R, L. M. & Konstantinidis, K. T. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Res.* **42,** e73 (2014).
36. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21,** 3674–3676 (2005).
37. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42,** D222–D230 (2014).

38. Glass, J. B. *et al.* Meta-omic signatures of microbial metal and nitrogen cycling in marine oxygen minimum zones. *Front. Microbiol.* **6,** 998 (2015).
39. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinforma. Oxf. Engl.* **28,** 3211–3217 (2012).
40. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinforma. Oxf. Engl.* **28,** 1420–1428 (2012).
41. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinforma. Oxf. Engl.* **22,** 2688–2690 (2006).
42. Dupont, C. L. *et al.* Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J.* **6,** 1186–1199 (2012).
43. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** (2011).
44. Konstantinidis, K. T. & Tiedje, J. M. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr. Opin. Microbiol.* **10,** 504–509 (2007).
45. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31,** 926–932 (2015).
46. Castelle, C. J. *et al.* Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nat. Commun.* **4,** (2013).
47. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
48. Berger, S. A., Krompass, D. & Stamatakis, A. Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Syst. Biol.* **60,** 291–302 (2011).
49. Orellana, L. H., Rodriguez-R, L. M. & Konstantinidis, K. T. ROCker: accurate detection and quantification of target genes in short-read metagenomic data sets by modeling sliding-window bitscores. *Nucleic Acids Res.* gkw900 (2016). doi:10.1093/nar/gkw900
50. Rho, M., Tang, H. & Ye, Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* **38,** e191 (2010).
51. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12,** 59–60 (2015).
52. Potter, L. C., Millington, P., Griffiths, L., Thomas, G. H. & Cole, J. A. Competition between Escherichia coli strains expressing either a periplasmic or a membrane-bound nitrate reductase: does Nap confer a selective advantage during nitrate-limited growth? *Biochem. J.* **344,** 77–84 (1999).
53. Khlebnikov, A. & Keasling, J. D. Effect of lacYExpression on Homogeneity of Induction from the Ptac and Ptrc Promoters by Natural and Synthetic Inducers. *Biotechnol. Prog.* **18,** 672–674 (2002).
54. Alberge, F. *et al.* Dynamic subcellular localization of a respiratory complex controls bacterial respiration. *eLife* **4,** e05357 (2015).
55. Hajaya, M. G. & Pavlostathis, S. G. Fate and effect of benzalkonium chlorides in a continuous-flow biological nitrogen removal system treating poultry processing wastewater. *Bioresour. Technol.* **118,** 73–81 (2012).

56. Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 5261–5266 (2002).
57. Thrash, J. C. *et al.* Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. *Sci. Rep.* **1,** (2011).
58. Luo, H. Evolutionary origin of a streamlined marine bacterioplankton lineage. *ISME J.* **9,** 1423–1433 (2015).
59. Rodríguez-Ezpeleta, N. & Embley, T. M. The SAR11 Group of Alpha-Proteobacteria Is Not Related to the Origin of Mitochondria. *PLoS ONE* **7,** e30520 (2012).
60. Viklund, J., Martijn, J., Ettema, T. J. G. & Andersson, S. G. E. Comparative and Phylogenomic Evidence That the Alphaproteobacterium HIMB59 Is Not a Member of the Oceanic SAR11 Clade. *PLoS ONE* **8,** e78858 (2013).
61. Konstantinidis, K. T. & DeLong, E. F. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J.* **2,** 1052–1065 (2008).
62. Takami, H. *et al.* A deeply branching thermophilic bacterium with an ancient acetyl-CoA pathway dominates a subsurface ecosystem. *PloS One* **7,** e30559 (2012).
63. Kuwahara, H. *et al.* Reduced genome of the thioautotrophic intracellular symbiont in a deep-sea clam, Calyptogena okutanii. *Curr. Biol. CB* **17,** 881–886 (2007).
64. Iobbi, C., Santini, C.-L., Bonnefoy, V. & Giordano, G. Biochemical and immunological evidence for a second nitrate reductase in Escherichia coli K12. *Eur. J. Biochem.* **168,** 451–459 (1987).
65. Iobbi-Nivol, C., Santini, C. L., Blasco, F. & Giordano, G. Purification and further characterization of the second nitrate reductase of Escherichia coli K12. *Eur. J. Biochem. FEBS* **188,** 679–687 (1990).
66. Philippot, L. Denitrifying genes in bacterial and Archaeal genomes. *Biochim. Biophys. Acta* **1577,** 355–376 (2002).
67. Martinez-Espinosa, R. M. *et al.* Look on the positive side! The orientation, identification and bioenergetics of 'Archaeal' membrane-bound nitrate reductases. *FEMS Microbiol. Lett.* **276,** 129–139 (2007).
68. Rothery, R. A., Workun, G. J. & Weiner, J. H. The prokaryotic complex iron–sulfur molybdoenzyme family. *Biochim. Biophys. Acta BBA - Biomembr.* **1778,** 1897–1929 (2008).
69. Yoshimatsu, K., Iwasaki, T. & Fujiwara, T. Sequence and electron paramagnetic resonance analyses of nitrate reductase NarGH from a denitrifying halophilic euryarchaeote Haloarcula marismortui. *FEBS Lett.* **516,** 145–150 (2002).
70. Lücker, S. *et al.* A Nitrospira metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 13479–13484 (2010).
71. Starkenburg, S. R. *et al.* Genome sequence of the chemolithoautotrophic nitrite-oxidizing bacterium Nitrobacter winogradskyi Nb-255. *Appl. Environ. Microbiol.* **72,** 2050–2063 (2006).
72. Sorokin, D. Y. *et al.* Nitrification expanded: discovery, physiology and genomics of a nitrite-oxidizing bacterium from the phylum Chloroflexi. *ISME J.* **6,** 2245–2256 (2012).

CONCLUSIONS AND RECOMMENDATIONS

SYNTHESIS AND RECOMMENDATIONS

This thesis presented a series of **advancements of laboratory protocols and associated bioinformatics techniques** for the study of complex microbial communities through omic techniques. We demonstrated the **integrated application** of those techniques and quantitative omics for assessing the functions and interactions of microbes at the levels of individual cells, populations of strains, and mixed species consortia. Our results showed that the integrated information from multiple levels of microbial assemblages is advantageous for better understanding, monitoring and manipulating mixed communities in engineered or natural ecosystems.

We first provided a framework for assessing and interpreting metatranscriptomics data (**chapter 2**), in order to quantify ***in situ*** **microbial gene activity** and identify key metabolic functions that are broadly applicable to environmental studies. We reported for the first time that the variability among technical replicates of metatranscriptomics datasets can be as high as between biological replicates. We offered a series of recommendations for future studies applying metatranscriptomics, including the minimization of sample handling times, improved mRNA extraction protocols, the use of replication, and the modeling of variance with fit to negative binomial distributions. Additionally, we showed that different RNA extraction protocols could bias the resulting metatranscriptomics datasets to a significant extent that can affect data interpretation. Depending on the targeted community, caution should be taken on choosing the most suited extraction protocols. For example organic extraction-based methods are likely to capture more of the high GC% content genomes in a sample.

Despite the technical considerations, given adequate replication, metatranscriptomic datasets can be used to assess the activity of even low abundance community members within an ecosystem, collectively referred to as the "**rare biosphere**". We used replicated transcriptomic data, generated together with accompanying time series metagenomic datasets, to characterize the expression profiles of a freshwater microbial community, and found two contrasting physiological adaptations among rare taxa. On one hand, transient rare members (i.e., members that were intermittently absent based on the time-series) typically showed disproportionally low gene expression rates. On the other hand, rare but consistent members presented disproportionally high expression rates, indicating that they are more active than the average cell in the community and thus, contribute to ecosystem functioning. The generalization and evaluation of these observations in other locations and habitats should be the focus of future work seeking to further understand and quantify the ecologic relevance of the rare biosphere and its role in ecosystem functioning and resilience.

Next, we presented a methodology based on 16S rRNA gene amplification and shotgun metagenome sequencing for **bacterial source tracking** in dynamic communities, and applied it to assess water quality of a freshwater habitat (Kalamas River, Greece) with documented anthropogenic contamination (**chapter 3**). This work demonstrated that the use of untargeted (shotgun) methods for contamination assessment is promising because they allow one to robustly detect contamination without a comprehensive list of biomarker organisms. However, additional standardization and testing is required in other locations and habitats. In the meanwhile, we recommend this approach being used alongside traditional methods such as coliform counts to establish a more robust, reproducible and sensitive procedure for water quality assessment.

We further present a novel technique for detecting **intra-species gene content differences *in situ,*** and we propose a new metric (core content index), for quantifying this level of micro-diversity in natural populations (**chapter 4**). We found that gene content diversity is significantly smaller for most natural populations compared to isolate collections of "named" species, with a few outlier populations exhibiting high gene content diversity. These findings suggest that a more stringent definition for species than the current definition is both more appropriate and plausible based on omics techniques.

The species definition is of cornerstone value in transporting and regulating microbes of clinical (pathogens) or environmental (biotechnology) importance. Our results also suggested that allelic diversity (measured by single nucleotide polymorphism frequencies) does not seem to correlate with gene content diversity in natural populations. This finding is consistent with the hypothesis that most of the gene content diversity within the population is neutral or nearly neutral, e.g. it does not confer a strong enough selective advantage to the cells that encode it in order for these cells to dominate the population and purge diversity. This observation further corroborates the conclusion that the identified populations are the key units (species) of microbial diversity. We speculate that several eco-evolutionary mechanisms may differentially affect both dimensions of diversity (gene content and allelic diversity), but it remains to be tested which of these mechanisms and to what extent promote or purge both allelic and gene content diversity.

The relevance of quantifying gene content in natural populations lies beyond the species definition. Our method can detect subtle differences in gene content of coexisting strains in the same system that might have implications for the **function and robustness** of microbial communities. We showcased the true value of the method for environmental studies (e.g., as demonstrated in **chapter 5**). For example, *Dehalococcoides mccartyi* populations with detected gene content differences, including reductive dehalogenases, coexist in highly enriched and specialized communities. This observed microdiversity might have implications in the robustness and reproducibility of dechlorination in mixed communities. For example, different strains might carry out distinct complementing functions in the system, such as consuming different chlorinated compounds and thus affecting the efficiency of dechlorination of mixed chlorinated compounds. Further experimentation using well-established consortia under different conditions (for example different electron donors), might reveal the underlying mechanism(s) for the maintenance of microdiversity and possibly further corroborate the findings reported here based on omics. Nevertheless, our work offers a framework for detecting and quantifying gene content diversity *in situ*, and is applicable to diverse systems, given the availability of high coverage metagenomic datasets.

We further applied quantitative metagenomic and metatranscriptomics analysis to identify key microbial functions two biological systems representing engineered and natural microbial consortia critical for environmental engineering studies. First, we

explore the functions and interactions within dechlorinating microbial consortia including *Dehalococcoides mccartyi (Dhc)* in order to identify the key mechanisms underlying the establishment of **robust dechlorinating communities** that can be used in bioremediation applications (**chapter 5**). We identified the major bacterial members in an actively dechlorinating community and quantified their relative contributions to metabolic functions that play important supporting roles on both dechlorinating efficiency and robustness. For example, at least five out of the 12 major community members are likely to support the growth of the major dechlorinator, *Dhc*, by providing essential vitamins such as biotin, thiamine and corrinoid precursors. A wide array of metabolic functions and symbiotic interactions were hypothesized in this work, serving as a blueprint to guide future efforts in the identification of conditions relevant for the faster and less labor-intensive establishment of robust dechlorinating consortia. Nevertheless, our results demonstrated a high degree of functional redundancy among community members, and gene expression data supported the hypothesis that the contribution of community members are overlapping and varying through time. For example, the highly enriched dechlorinating community studied here consists of multiple dechlorinators that compete for the same substrates. Non-dechlorinating community members seem to be collectively supporting the various nutritional requirements of the dechlorinating organisms, and multiple strains of the same species (such as *Dhc*, described above) coexist in the system. Those observations taken together point out to the importance of mixed consortia *versus* axenic cultures in successful robust dechlorination activity: different species carry out the same function (i.e., TCE dechlorination) and the strains of the same species could potentially carry out different functions (i.e., gene content diversity in *Dhc*). Thus, both functional redundancy and labor division could influence the community activity, and explain why mixed consortia are more efficient than highly specialized axenic cultures.

Finally, we applied integrated omic techniques to identify key metabolic adaptations for the ubiquitous SAR11 bacteria, within the anoxic water masses formed in oceanic Oxygen Minimum Zones (OMZs) (**chapter 6**). Studying the microbial activities within the OMZ ecosystems is crucial for understanding and predicting how those systems affect the nutrient cycles on a global scale. OMZs are considered “hot spots” of nitrogen loss, accounting for up to 50% of total N lost from the ocean to the atmosphere as N_2 or N_2O gas. The predominant microbial activities leading to the nitrogen loss are

denitrification and anammox in OMZs, which both depend on the first step of nitrate reduction. SAR11 organisms were traditionally considered obligate aerobic heterotrophic microbes, thus their high abundance within the OMZ anoxic waters remained a puzzling observation. We identified for the first time respiratory nitrate reductases (Nar) in OMZ SAR11 genomes, which represent novel clades specific to the anoxic niche and constitute a major fraction (up to 30%) of the OMZ community. We also showed that the majority of the genes mediating nitrate reduction in the OMZ are transcribed by SAR11. Our results implicate SAR11 as a primary contributor to these pathways and thus, establish it as a model organism for future studies of the nutrient cycling in OMZs and possibly nitrogen retention in the system. Those results significantly advance our understanding of the microbial players controlling nitrogen (N) cycling in the ocean. Moreover, this work showcases the integration of omics techniques with functional characterization of key OMZ resident populations to link functions to otherwise inaccessible organisms (i.e., uncultivated). Building upon this work, future comparative analysis of SAR11 single cell genomes from various oceanic regions can unravel the adaptive mechanisms that allow SAR11 cells to occupy diverse environmental niches, and provide insights into the genomic versatility of this globally dominant organism as well as the biotic and abiotic controls of its activity.

APPENDIX A
SUPPLEMENTARY MATERIAL FOR CHAPTER 2

SECTION A.1: DETAILED PROTOCOLS

A.1.1 Nucleic acid extractions

Enzymatic lysis plus OP (organic extraction protocol) RNA isolation: The organic extraction method for RNA was performed as described previously [1] with minor modifications. In brief, lysis buffer (50 mM Tris-HCl, 40 mM EDTA, and 0.75 M sucrose) was added to the filters, followed by addition of 1 mg/ml lysozyme and incubation at 37°C for 30 min. The lysates were subsequently incubated with 1% SDS and 10 mg/ml proteinase K, for 2 h at 55°C in a rotating hybridization oven. RNA was extracted from lysate with acid phenol and chloroform, and isolated using filter columns from mirVANA RNA isolation KIT (Ambion), washed twice following the manufacturer's instructions and eluted in TE buffer.

Bead-beating lysis plus kit-based protocol BP (bead-beating protocol) RNA isolation: RNA was extracted using a modified RNeasy Kit (Qiagen) protocol [2]. Samples were first vortexed for 10 min with RNase-free beads from the Mo-Bio RNA PowerSoil kit (Carlsbad, CA). Following centrifugation for 5 min at 5,000 × *g*, the supernatant was transferred to a new tube. One volume of 70% ethanol was added to the lysate, which was drawn up through a 22-gauge needle several (~5) times to shear genomic DNA. RNA extraction then continued with the RNeasy Mini kit according to the manufacturer's instructions.

Following extraction, all RNA samples were treated with DNase using the TURBO DNA-*free* kit (Ambion, Austin, TX) with 2µl of TURBO-DNase per 100 µl of

reaction volume, which according to our experience removes DNA contamination from similar samples. In all but one of the cDNA libraries, the rRNA depletion was performed with a combination of enzymatic rRNA degradation (mRNA-only Prokaryotic mRNA Isolation Kit, Epicenter) followed by two treatments of rRNA subtractive hybridization (MICROB*Express*, Ambion). The subtractive hybridization-based Ribo-Zero rRNA removal kit (Epicentre) was applied in only one RNA prep (OP5B) as an alternative to the two rounds of MICROB*Express* treatment, in order to test this newly commercially available KIT at that time. Enriched mRNA samples were linearly amplified using the MessageAmp II-Bacteria Kit (Ambion), reverse transcribed with random hexamers using the Universal RiboClone cDNA Synthesis System (Promega, Madison, WI) and purified with the MinElute DNA clean-up kit (Qiagen).

DNA was extracted from the sterivex filters as previously reported [3], using a similar cell lysis and organic extraction method as described above with the following modifications for DNA isolation: RNAse was added to the lysis buffer at a concentration of 200 µg/ml, phenol (pH 8.5) and chloroform were used for the extraction of DNA from the lysates, and DNA was precipitated with cold ethanol and subsequently washed twice with 70% ethanol.

All nine cDNA and seven DNA libraries were prepared in the same manner according to Illumina standard protocols, and were sequenced on the Illumina GA II sequencer from both ends (paired end, 2x100; 300-500 bp long inserts) at Emory University Genome Center.

A.1.2 Quality filtering of sequences and rRNA identification

cDNA and DNA sequences were trimmed using the following criteria: Probability of error cutoff of 1% at both ends, minimum length of 50 bp, and removal of N-containing sequences. At this step, two out of the nine cDNA datasets (BP3B, OP2) were excluded from further analysis due to large number of low quality reads (>70% of the total). cDNA datasets were further filtered to remove poly-A tails and chimeric sequences as previously described [4]. Non-coding RNA sequences were identified by a Blastn search [5] (default settings; cut-off: bit score >50) against the SILVA small and large subunit rRNA

(http:/www.arb-silva.de) and RFam (http://rfam.janelia.org) databases. Only coupled reads that passed the above quality criteria were maintained for further analysis.

A.1.3 Evaluation of reproducibility among replicates

Pairwise comparisons of cDNA datasets: The datasets were compared pairwise at 3 different levels:

(i) Relative Gene expressions: cDNA reads were mapped to the predicted genes and the counts were scaled on each dataset proportionally to the size of the smallest dataset and rounded. For example, the abundance of a gene *i* in dataset A (a_i^A) with c_i^A reads mapped (raw count of *i*), was defined as:

$$a_i^A = \left\lfloor \frac{c_i^A \cdot \sum_{j \in S} c_j^S}{\sum_{k \in A} c_k^A} \right\rfloor \qquad \textbf{(1)}$$

Where S is the smallest dataset. The resulting values where rounded down to account for the different detection limits in datasets of different size. The relative gene expressions were used to compare the datasets using both the Pearson's correlation distance ([1-R]/2) in log-log space (excluding zeroes) and the Bray-Curtis dissimilarity index using the R package ecodist [6].

(ii) Relative abundance of ORF clusters: Since around 50% of the cDNA reads were mapping to the metagenomic assembly, we adopted a strategy to directly compare clustered ORFs predicted from cDNA reads, aiming to represent protein families [7]. cDNA reads were partitioned into clusters of ORFs (see below *ORF prediction and clustering*) in order to compare the datasets without database biases, derived from the need of assembly or annotations. Counts were scaled down as above (equation 1), and clusters recruiting less than 10 reads across datasets were excluded. Pearson's correlation distances and Bray-Curtis dissimilarity indices were calculated as described above.

(iii) K-mer compositions: The last level of comparisons included all the cDNA reads by contrasting the k-mer distributions of the reads from each dataset, described below in detail (*K-mer composition analyses*). Briefly, the k-mer distributions allowed for the identification of reads "typical" in one dataset, versus another for

each pairwise comparison, using a Log-likelihood score. Hellinger distances were calculated between the distributions of those scores, aiming to quantify the compositional dissimilarity of datasets at the read level.

All distances were visualized by clustering the datasets using the Ward method (built-in R function hclust) as shown in Figure 1, main text. Additionally the distributions of fold changes in relative gene expressions among pairs were compared between technical, biological or intra-protocol replicates (Figure B2).

ORF prediction and clustering: Open Reading Frames (ORFs) were predicted in all six reading frames from the cDNA reads using getorf [8]. In order to reduce the redundancy of each dataset while accounting for sequencing errors, ORFs longer than 30 amino acids were clustered at 98% nucleotide sequence identity and alignment length of at least 80% of the length of the shortest sequence using cd-hit [9] (98%ORFs). The non-redundant 98%ORFs from each dataset were subsequently clustered at 60% amino acid sequence identity and 80% length of the shortest sequence among all cDNA datasets to provide a list of non-redundant protein families. The occurrence and relative abundance of ORF clusters at both clustering levels (98 and 60%) were compared across all cDNA datasets. For functionally annotating ORFs, Blastp (only best hit considered, cutoff > 60 bits score) was used to compare representative sequences from the non-redundant ORF clusters (98% level) from each library against the GOS peptides [10].

K-mer composition analyses: The k-mer composition of the reads in each cDNA dataset was analyzed to achieve two goals: first, to distinguish and quantify potential sequencing errors and artifacts from low abundant transcripts and second, to compare the distributions of k-mers in a pairwise manner for all cDNA datasets to assess the sequence similarity and differences of the datasets at the read level.

For the first level of comparisons, the relative abundance of a transcript within a cDNA dataset was compared against its relative abundance across all replicated cDNA datasets. The k-mer mode of each read within a dataset was used as a proxy for the former, and the size of the ORF cluster that the read was assigned to was used as a proxy for the latter (ORF clusters were built using all datasets in this case). The distribution of 21-mers was calculated for each dataset using Jellyfish [11]. The k-mer

mode of each read was calculated as the mode of the number of occurrences of all 21-mers identified on that read within the dataset of origin. Reads representing rare transcripts within a dataset were flagged as k-mer mode one; the rest were flagged as k-mer mode greater than one. These Boolean values were compared against the ORF cluster size. For each read, the ORF cluster size is defined as the size of the cluster containing the ORF derived from the read. For those reads with multiple predicted ORFs at different reading frames only the largest ORF cluster was considered. Reads without ORFs longer than 90 nucleotides were assumed to have cluster size of zero (likely representing non-coding sequences or sequencing errors).

For the second level of comparisons, the following analysis was performed: For any pair of datasets, the log-likelihood ratio ($LLratio_{A:B}$) for a given read (r) being compositionally more typical in one dataset (A) versus the other (B) was estimated using the 21-mer distributions of the datasets under comparison (eq. 2).

$$LLratio_{A:B}(r) = \sum_{k \in k_r} \log\left(\frac{P(k \mid A)}{P(k \mid B)}\right) \qquad \textbf{(2)}$$

Where k_r is the collection of k-mers composing the read r, and the probability of the k-mer k given the dataset A is defined in equation 3:

$$P(k \mid A) = \begin{cases} \dfrac{N_k^A + 1}{\sum_{l \in K_A} (N_l^A + 1)}, \forall N_k^A > 1 \\ \dfrac{1}{\sum_{l \in K_A} (N_l^A + 1)}, \forall N_k^A \leq 1 \end{cases} \qquad \textbf{(3)}$$

Where N^i_k is the number of times the k-mer k occurs in the dataset i, and K_i is the non-redundant collection of k-mers in the dataset i (excluding k-mers occurring only once). Next, the distributions of $LLratio_{A:B}$ for reads from dataset A and B were obtained discarding reads with k-mer mode one, and these distributions were used to define the significance of the read being compositionally more typical of one dataset versus the other. A read from dataset A was considered significantly more typical of A (with maximum error type I α=0.01) if the log-likelihood ratio was higher than the 99-percentile of the $LLratio_{A:B}$ distribution of dataset B. Conversely, a read from dataset B was

considered significantly more typical of *B* (with α=0.01) if the log-likelihood ratio was lower than the 1-percentile of the $LLratio_{A:B}$ distribution of dataset *A*. These thresholds were used to select all the compositionally different reads from each dataset of a pairwise comparison. The difference in G+C% content of these reads was assessed using a two-tailed t-test between the two datasets compared (α=0.01).

For each pair-wise comparison, the difference between the compositions of the datasets was estimated using the Hellinger distance (eq. 4) between the distributions of log-likelihood ratios described above.

$$H(A,B) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=-\infty}^{+\infty} \left(\sqrt{f_i^A} - \sqrt{f_i^B} \right)} \qquad \textbf{(4)}$$

Where f_i^A represents the frequency of the value *i* in the $LLratio_{A:B}$ distribution of the dataset *A*. Note that *H(A,B) = H(B,A)*.

For all possible pair-wise comparisons, the Hellinger distance as well as the average G+C% content of sets of compositionally different reads were computed. Each pair-wise comparison was categorized as inter-protocol (OP:BP and BP:OP) or intra-protocol (OP:OP and BP:BP), and the Hellinger distances and average G+C% contents were compared between categories using two-tailed t-tests (α=0.05).

Assessment of over-dispersion across datasets: We evaluated the fit of several distribution models to our data, as proposed elsewhere [12]. Over-dispersion with respect to the Poisson distribution of read counts was graphically explored based on the relationship between mean and variance of gene expressions and the Poisson, over-dispersed Poisson (OD-Poisson), and negative binomial (NB) distributions [12]. For the Poisson distribution, no parameter estimation is necessary (*variance = mean*). In the OD-Poisson distribution (*variance = k × mean*), the *k* parameter was estimated as the slope of the linear regression between mean and variance. For the NB distribution (*variance = mean + ϕ × mean*^{2}), one global *ϕ* parameter was estimated per comparison, using the edgeR package [13]. Over-dispersion was quantitatively assessed by comparing the estimated *ϕ* parameters for the NB distribution, exploiting the relationship on (eq. 5), according to which larger values of *ϕ* imply closer resemblance to Poisson (*i.e.*, negligible over-dispersion).

$$Poisson(\lambda) = \lim_{\varphi \to \infty} \left[NB\left(\varphi, \frac{\lambda}{\lambda + \varphi}\right) \right] \quad \textbf{(5)}$$

A.1.4 Characterization of population genomes

Population genomes were recovered using a similar approach to that previously described [14]. Briefly, a graph was built where nodes represented contigs and edges represented high correlation in the abundance of individual contigs across time points (R^2 > 0.85) or high correlation in tetra-nucleotide composition (distribution of tetra-nucleotides as Z-scores, R^2 > 0.9). The graph was divided in clusters using the silhouette method [15] implemented in the R package fpc, the network was visualized with neo4j, and contigs were partitioned after trimming edges with Pearson's coverage correlation <0.95, tetranucleotide z-score correlation < 0.9, and with less that 3 paired end connections. Reads mapping to the contigs on each partition were isolated and re-assembled using Velvet [16]. This process was iterated until no further increase in N50 was observed. Quality assessment of the resulting population genome sequences was based on evaluation of coverage evenness in each metagenome, identification of 103 single copy universal genes [14] with HMMer searches [17], and phylogenetic reconstruction of 30 marker genes identified by AMPHORA2 (Wu & Scott 2012), with RAxML v7.4.2 (-f a –N 100 –m PROTGAMMAGTR) [18]. Population genomes were considered for further analysis when at least 75 single copy universal markers were identified and the 30 different phylogenetic reconstructions were consistent.

Genes in population genome sequences were predicted with GeneMark (Lukashin & Borodovsky 1998) using the phylogenetically closest available model. Functional annotations were performed using the same pipeline as described for the metagenomic genes with the addition of annotations from the Blast2GO pipeline (Götz et al. 2008) as well as KEGG Orthology (Kanehisa et al. 2012) using blastp (only best hit considered, score cut-off ≥ 90 bits). Potentially complete pathways were identified using MinPath (Ye & Doak 2009) and manually curated based on the BRITE reference hierarchies, excluding the human diseases and organismal systems categories.

For each genome, gene expression values were calculated from the average normalized cDNA counts with the DEseq package. Counts were further normalized for the gene length (reads per kb, rpk) to identify those genes with the highest relative

expression within each genome. To identify significantly overrepresented functional categories among the transcriptomes of the recovered genomes, a matrix of averaged normalized cDNA counts was built, where rows represented the population genomes and columns the functional categories identified. For each cell of the matrix, a 2x2 contingency table was constructed and a one-sided Fischer exact test was applied. Odd ratios producing p-values below 1E-15 were selected as functional categories significantly enriched in the transcriptomic profile of the given genome when compared to all others. The relative expressions of nutrient transporters were estimated by adding the expression values of genes identified as dissolved organic carbon transporters based on their COG annotations using a similar approach as in [2]. Relative expressions of identified essential inorganic nutrient transporters were calculated in a similar manner, based on keyword searches in the annotation descriptions and GO terms from the Blast2GO pipeline (Figure B9).

SECTION A.2: SUPPORTING RESULTS AND DISCUSSION

A.2.1 Sequence noise filtering and rare sequences

A common step in high-throughput sequencing data analyses is filtering out possibly spurious sequences based on their abundance measured, for instance, by the k-mer mode [19–21]. With this filter, sequences with a k-mer mode of one within a dataset (singletons) are assumed to be either sequencing errors, artifacts, or rare transcripts that do not allow statistical resolution. This step potentially represents a double advantage for downstream analyses. First, filtering out spurious sequences reduces the impact of noise, allowing for more meaningful differences between treatments to be detected. Second, it largely reduces the number of sequences, improving computational tractability. If reads with a k-mer mode of one were indeed spurious sequences, these reads would have a higher probability of being singletons or not detectable across replicated datasets, and consequently would be assigned to smaller protein clusters across technical and biological replicates relative to reads with k-mer mode greater than one.

In order to assess the suitability of the k-mer method for sequence noise reduction, the extent to which the read k-mer mode within each dataset reflects the abundance of the ORF cluster that the read is assigned to was evaluated (Supporting Methods). ORF clusters were defined as sets of reads (from all seven metatranscriptomic datasets) with at least 60% amino acid identity over 80% or more of their length, accounting for potential sequencing errors and protocol-specific G+C% biases. ORF cluster abundance was directly compared with the k-mer mode of the reads. As expected, sequences with k-mer mode greater than one rarely included singletons across datasets and were typically assigned to large ORF clusters (Figure B2). About 70% of the sequences with k-mer mode one were not singletons across datasets, and about 50% of them were assigned to clusters with 10 or more members. The functional annotation of the reads also showed that sequences with k-mer mode greater than one did not present significantly more annotated functions than those with k-mer mode of one (Figure B2, inset). Therefore, k-mer mode of one did not reflect well singletons or reads assigned to small clusters, indicating that the corresponding reads may or may not be spurious sequences, and should not be excluded. In other words, rarity of a transcript within a dataset should not be used as a quality filtering; instead deep sequencing and replication, which provide statistical resolution for such rare transcripts, should be employed.

A.2.2 Overdispersion across cDNA datasets

Directly assessing the level of variation is critical for comparative transcriptomic studies because when over-dispersion is large, assumptions of some statistical methods are violated. More specifically, statistical modeling of digital expression values is typically accomplished using the Poison distribution of probability [22]. In over-dispersed data, however, the parameter of the distribution (λ) cannot be assumed to be constant (typically suitable for technical replication); instead it can be modeled using a gamma distribution, resulting in digital expression values being modeled with a Negative Binomial distribution with two parameters (ϕ and p; [13,23]). When ϕ is large enough the Negative binomial distribution is indistinguishable from the Poisson distribution (with parameter $\lambda=\phi p/(1-p)$), so a larger estimated value of ϕ implies smaller over-dispersion and a better fit to a Poisson distribution with constant λ. The OP5A and OP5B datasets

were prepared with different mRNA enrichment protocols but displayed the smallest dispersion in this study and fit well a Poisson distribution ($\phi \approx 32.1$). Subtractive hybridization treatments (involved in both preparations) have been shown to be highly reproducible with high integrity RNA [24], and the variation introduces doesn't seem to exceed the technical variation, introduced during the identical treatment but independent preparations of aliquots from the same sample. Indeed OP5A and OP5B could be considered "technical replicates" based on the low variation observed. However, all other tested collections of replicates displayed a poor fit to a Poisson distribution (ϕ between 2.3 and 4.3); hence, biological and technical variability (*i.e,* OP4A, OP4B) could not always be differentiated on the basis of over-dispersion. In such cases, a negative binomial distribution should be preferred [see also [23] for a method for negative binomial fit without replication]. These results highlighted the importance of testing over-dispersion on comparative transcriptomic studies whenever possible (*e.g.*, when replication is available for all treatments).

A.2.3 Extraction protocol biases and implications

When comparing pair-wise similarities in gene expression between all datasets, it became evident that variability between extraction protocols exceeds variability from both technical and biological replicates. Not only did datasets from the same extraction protocol clustered together (Figure 1), but also the largest over-dispersion was observed when datasets from different protocols were treated as replicates to fit a Negative Binomial distribution. Moreover, the variation observed between extraction protocols largely exceeded that of the pair of datasets OP5A/OP5B derived from different rRNA subtraction protocols (Figure 1, Figure B3). In order to explore possible causes of the systematically higher variation between the different extraction protocols, k-mer abundance distributions were computed and compared across libraries. Advantages of this comparison are that it explores entire datasets (not only annotated portions) and displays differences at the read level, independently of potential biased binning. These advantages are important because about three quarters of the reads cannot be mapped to the reference genes (Figure B5) and binning of reads (*e.g.*, by putative homology, function, or taxonomy) can introduce biases in sequence composition. When the k-mer distributions were compared, datasets did not cluster by protocol, indicating that

compositional bias alone is not sufficient to distinguish the protocols. However, when identifying reads more typical of OP datasets versus the BP datasets (pair-wise dataset comparisons; p-value < 0.01, log-likelihood of k-mers, see Supporting Methods) based on their k-mer content, there was a clear G+C% difference observed. It was found that the G+C% content of the reads more typical of the BP reads was always significantly higher than that of OP reads (p-values < 1e-16, two-tailed t-test), while no significant difference was found when datasets originating from the same protocol were compared (p-value = 0.13, t-test, Figure 1). Notice that using this approach, among all pairwise comparisons on average 17% (ranging from 4-34%) of the total reads from each dataset were deemed as typical in one versus the other dataset.

This G+C% bias was also evident when comparing the datasets at the functional or taxonomic annotations level. When the datasets from the two protocols were treated as different treatments, only few genes and taxa were found to be differentially represented, mostly related to viral proteins and genomes of high or low G+C% (Table S3). In most comparative transcriptomic studies such differences (biases) are of reduced importance because they are systematically introduced in all treatments (same protocol), but they may lead to misleading interpretations in descriptive transcriptomic studies. Moreover, it has been recently proposed that environmental sequencing data should be more frequently subjected to meta-analysis, as opposed to project-specific analyses (*e.g.*, Tartar, Wheeler et al. 2009; Delmont, Malandain et al. 2011), and our work shows that differences in extraction methods can introduce noisy, yet apparently relevant, differences in expression profiles

SECTION A.3: SUPPLEMENTARY FIGURES AND TABLES

Table A1. Sample information and sequencing statistics for the cDNA and DNA datasets used in this study.

			Sampling		Dataset size (Gb)		Physicochemical Parameters				
	Dataset	# of quality filtered reads [c]	SRA accession numbers	Date	Time	Raw Reads	Qual-filtered	T (°C)	pH	Turbidity (NTU)	Dis. solids (g/L)	Diss O_2 (mg/L)
Metatranscriptomes	OP2 [b]		SRR949987	06/07/10	16.12	11.21	5.14					
	OP5A	43,342,026	SRR949847	06/07/10	16.55	14.15	10.38					
	OP5B	7,424,413	SRR949897	06/07/10	16.55	1.94	1.33					
	OP4A	2,660,964	SRR949535	06/07/10	16.45	0.8	0.58					
	OP4B	4,093,349	SRR949536	06/07/10	16.45	1.98	1.10					
	BP2	1,314,768	SRR949953	06/07/10	16.12	0.95	0.87	30.5	7.5	5.2	0.027	8.93
	BP3A	1,918,014	SRR949314	06/07/10	16.30	1.27	0.77					
	BP3B [b]		SRR949315	06/07/10	16.30	2.41	0.71					
	BP5	1,556,845	SRR949950	06/07/10	16.55	1.02	0.67					
Metagenomes	MTG1	57,973,824	SRR948284	06/07/10	16.17	9.11	8.16					
	MTG2	59,198,606	SRR948155	06/07/10	16.12	9.91	8.6					
	AUG09 [a]	24,155,706	SRR096386	26/08/09	12.10	6.48	4.5	NA	NA	NA	NA	NA
	NOV09 [a]	25,252,022	SRR096389	08/11/09	13.25	6.99	4.55	19.2	6.78	3.2	0.03	7.33
	SEP10	31,605,662	SRR947737	10/09/10	16:16	8.53	5.83	28.8	7.6	5.3	0.03	7.52

Table A1 continued

NOV10	36,321,612	SRR948334	14/11/10	16:33	9.47	6.78	17.9	5.73	4.1	0.03	5.54
JAN11	33,803,022	SRR948448	29/01/11	16:09	10.5	6.38	7.8	6.56	5.4	0.03	4.16

Sequencing was performed on the Illumina GA II platform using a paired-end read strategy (2X100).

[a] These datasets were previously described in [3]. All other datasets were sequenced as part of this study.

[b] Datasets were excluded from further analysis due to high fraction of low quality reads.

[c] Number of putative mRNA reads for metatranscriptomes

Table A2. RNA yield at each step of the metatranscriptomic protocols.

Dataset ID	Filter ID	Part of filter used	Extraction Method	Total RNA yield (ng of RNA)	Post mRNA enrichment yield (ng of mRNA)	Post amplification yield with 100 ng mRNA input (µg of aRNA[a])	RT yield[b] with 6 µg of aRNA[a] (µg of cDNA)	Sequencing yield (Gb)
OP2	2	½	OP	951	480	26	3.2	11.21
	3	½	OP	765	270	17	4	NA
OP4A	4	¼	OP	240	120	22	2.4	0.58
OP4B	4	¼	OP	285	108	13	2.7	1.1
OP5A	5	¼	OP	930	512	9	3	14.15
OP5B	5	¼	OP	375	< 125	20	3	1.33
BP2	2	½	BP	1140	690	8	3	0.95
BP3A	3	¼	BP	496	243	15	3.4	1.27
BP3B	3	¼	BP	270	184	12	3.3	2.41
	4	½	BP	525	132	50	5	NA
BP5	5	½	BP	814	527	12	3.8	1.02

Note that no systematic differences in RNA yields were observed between OP and BP protocols after the extraction, amplification, or reverse transcription steps. However, OP datasets resulted in generally larger sequencing yields. Two samples (2nd and 10th) were not sequenced, and hence no dataset IDs are associated with those.

[a] Amplified antisense RNA (aRNA).

[b] Reverse transcription yield (RT yield).

Table A3. Taxa that showed statistically significant different expression levels between the two extraction protocols.

Enriched in	Species	Genome size (Mbp)	G+C%	Cell wall
BP	*Roseobacter denitrificans*	4.33	59	G-
BP	*Roseobacter litoralis*	4.67	57.3	G-
BP	*Heliobacterium modesticaldum*	3.1	57.0	G+
BP	*Ruegeria sp. TM1040*	4.15	60.1	G-
BP	*Ruegeria pomeroyi*	4.11	64.2	G-
OP	*Parachlamydia acanthamoebae*	5.8	35.5	G+
OP	*Bacillus_weihenstephanensis*	5.26	37	G+
	Genus			
BP	*Roseobacter*	4.35±0.4	58±1.9	G-
BP	*Candidatus Puniceispirillum*	2.75	48.9	G-
BP	*Heliobacterium*	3.08	57.0	G+
BP	*Ruegeria*	4.2±0.36	60.6±2.3	G-
OP	*Candidatus Liberibacter*	1.3±0.12	35.9±0.72	G-
	Phylum			
BP	*Cyanobacteria*	5.1±2.2	43.2±8.8	G-
OP	*Crenarchaeota*	1.9±0.4	45.5±9.3	
OP	*Chlamydiae*	1.33±0.5	40.2±1.1	G-
OP	*Euryarchaeota*	2.44±0.7	44.3±12.3	
OP	*Nanoarchaeota*	0.49	31.6	

The table shows the differentially expressed taxa at the species, genus and phylum levels that were significantly enriched in OP or BP datasets (p-value adjusted ≤ 0.01, negative binomial test). Contigs were taxonomically classified by the MyTaxa pipeline (see Materials and Methods). Average genome sizes and G+C% content, including standard deviations when multiple genomes were available, were calculated from the available sequenced genomes in NCBI assigned to the same taxon. Note that low G+C% content genomes were typically enriched in OP datasets and high G+C% content genomes in BP.

Table A4. Functional categories enriched in OP datasets.

Enriched GO categories on OP datasets	
GO:0003677	DNA binding
GO:0003910	DNA ligase (ATP) activity
GO:0004519	endonuclease activity
GO:0004748	ribonucleoside-diphosphate reductase activity, thioredoxin disulfide as acceptor
GO:0005198	structural molecule activity
GO:0005524	ATP binding
GO:0005737	cytoplasm
GO:0006260	DNA replication
GO:0009186	deoxyribonucleoside diphosphate metabolic process
GO:0016032	viral reproduction
GO:0016730	oxidoreductase activity, acting on iron-sulfur proteins as donors
GO:0016853	isomerase activity
GO:0019028	viral capsid
GO:0019048	ribonucleoside-diphosphate reductase activity, thioredoxin disulfide as acceptor
GO:0019073	viral DNA genome packaging
GO:0019685	photosynthesis, dark reaction
GO:0030430	host cell cytoplasm
GO:0030494	bacteriochlorophyll biosynthetic process
GO:0044419	interspecies interaction between organisms
GO:0046797	viral procapsid maturation
GO:0046872	metal ion binding
GO:0046914	transition metal ion binding
GO:0050577	GDP-L-fucose synthase activity
GO:0050662	coenzyme binding

Gene expression values were binned into GO annotations and OP/BP libraries were compared using the DESeq statistical package to identify GO terms overepresented in OP datasets (p-value adjusted ≤ 0.01, negative binomial test).

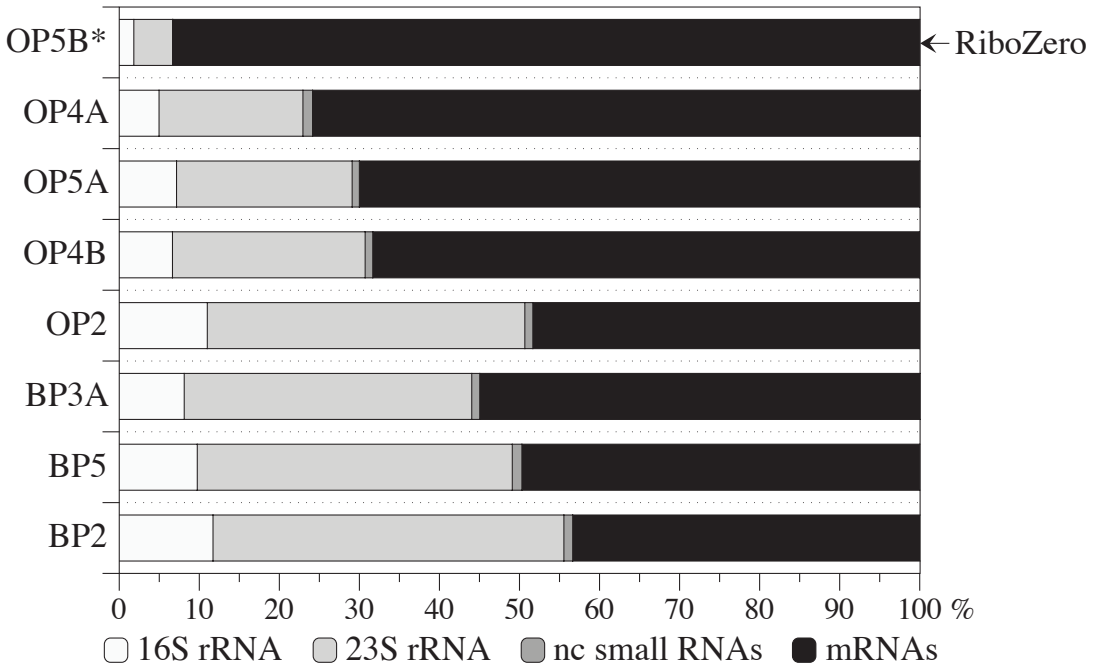


Figure B1: **Performance of rRNA depletion protocols.** Percentage of sequences that were identified to encode 16S and 23S rRNA genes, other non-coding small RNAs or putative coding mRNAs in the sequenced libraries (figure key). All preparations, except OP5B (Ribo-Zero protocol), included mRNA-Only and MicrobeExpress treatments and showed various levels of non-rRNA sequence enrichments, ranging from 40-70%. The Ribo-Zero treatment showed the highest rRNA depletion.

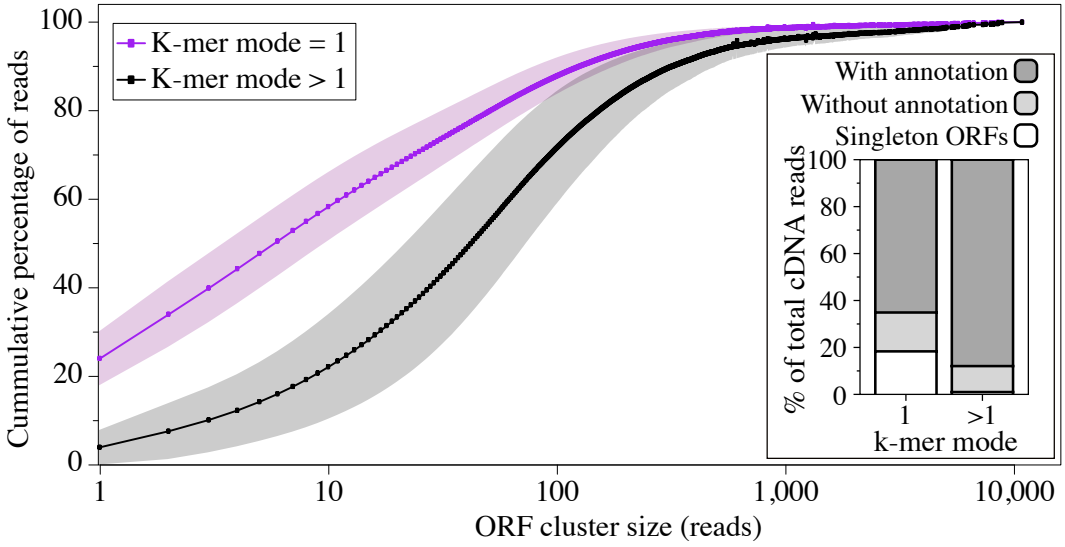


Figure B2. **K-mer abundance analysis of metatranscriptomic sequences.** Sequence reads were assigned to ORF clusters as described in the Supporting Methods section and the graph shows the cumulative percentage of reads assigned to clusters with k-mer mode one (representing singleton reads within the dataset of origin; purple line) or reads with k-mer mode greater than one (representing not singleton reads; black line). The shaded areas represent the average ± one standard deviation based on all replicate cDNA libraries. **Inset**: Functional annotation of reads with k-mer mode of one or greater. Reads that were represented more than once within ORF clusters based on all cDNA datasets (not singletons) were categorized into annotated (if the representative ORF had a match in GOS protein clusters, SwissProt, or closed prokaryotic genomes) or non-annotated (no match); singleton ORFs were not annotated. Note that on average, 50% of the sequences with k-mer mode one were members of ORF clusters with ≥10 reads across replicates and more than 60% of those are functionally annotated, indicating that they represent real sequences and thus, should not be filtered out (although the average percentage of singleton ORFs was higher in reads with k-mer mode one relative to those with k-mer mode greater than one).

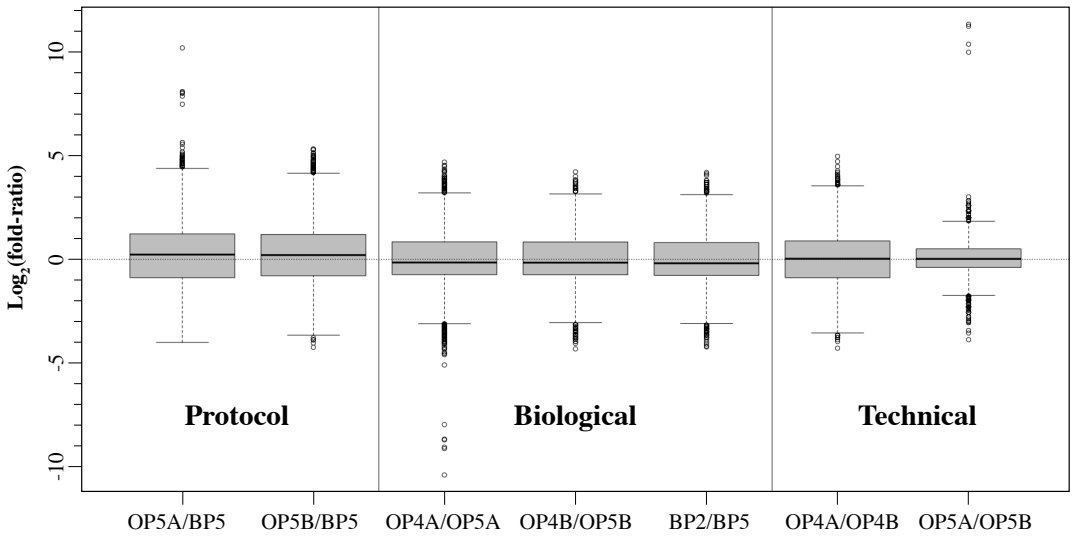


Figure B3. **Comparison of inter and intra-protocol variability among technical and biological replicates.** Box-plots represent the distribution of gene expression fold changes (log_2 ratios of normalized gene counts for each pair-wise comparison). The first two comparisons represent protocol comparisons (same biological sample) followed by three pair-wise comparisons of biological replicates and two technical replicates. Note that variability is typically higher in inter-protocol than intra-protocol comparisons (technical or biological). The two technical replicate comparisons (right) show similar variation to the biological replicates (center). Calculations were performed only for genes present in both datasets examined in each case.

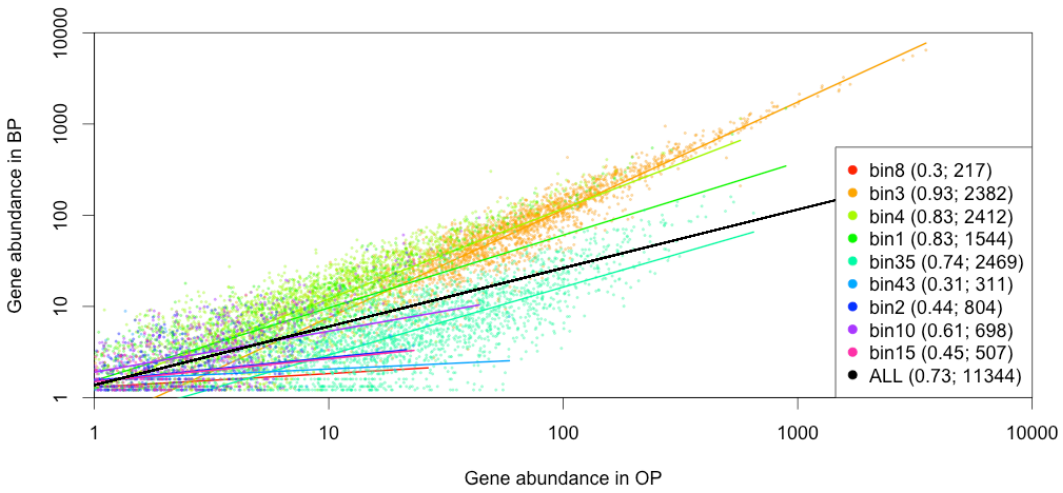


Figure B4. **Correlation of gene transcript abundance between OP and BP datasets for each population genome (bin) recovered.** The lines represent the linear regression in log-log scale of the gene abundance values for each bin. The number of genes used in each comparison and the Pearson correlation coefficients of the analysis are shown (figure legend) for each comparison together with the number of genes identified in both datasets and used to calculate the correlation values. Gene abundance values were calculated as average normalized counts using the DESeq package (four OP datasets, and 3 BP datasets). The population genomes were recovered from the available time series metagenomes, using different binning strategies. Nevertheless each bin can be assumed to represent a population of almost identical cells. Note that the correlation in abundance values is 0.73 when all genes are considered. However, the correlations are slightly higher when each bin is considered separately (when the number of genes is large). These results might indicate different cell lysis efficiencies for different organisms between the two protocols, since the correlations are high but the slopes differ for each population genome.

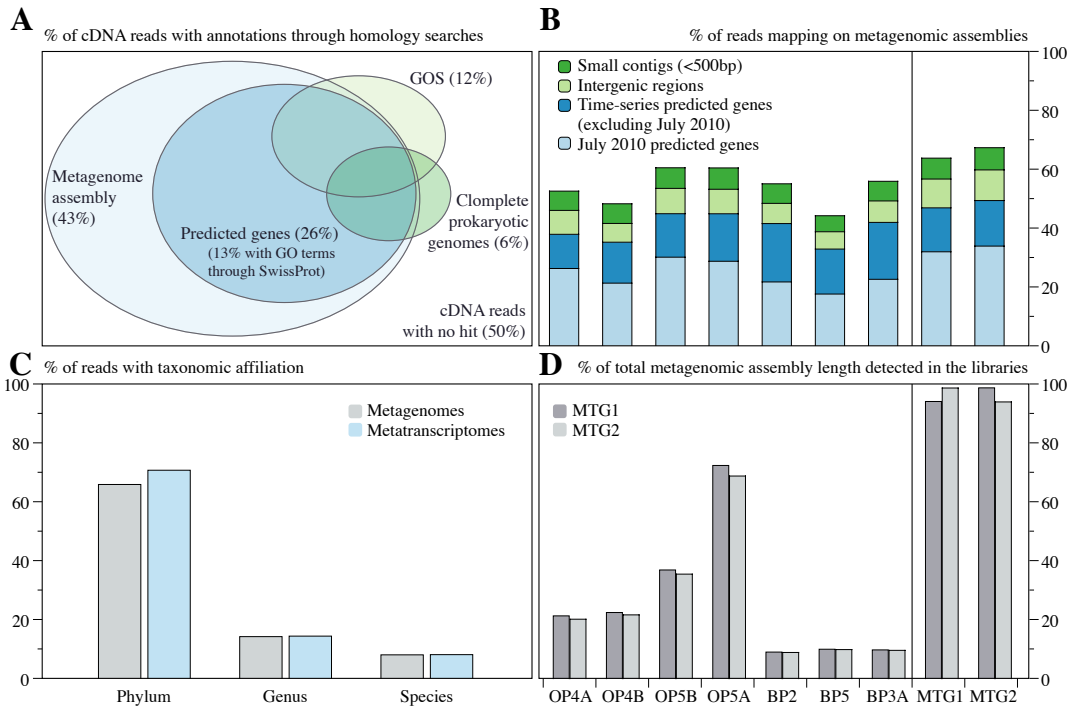


Figure B5. Percentage of the datasets annotated with different strategies. (A) Percentage of non-ribosomal cDNA sequences from all metatranscriptomic datasets that were classified based on homology searches against selected databases and protein families. Only 43% of the total cDNA sequences mapped on assemblies from companion DNA samples and less than half of those sequences were collectively mapped on GOS (Global Ocean Sampling) protein sequences and available closed genomes. (B) Percentage of the metatranscriptomic datasets mapped to different subsets of the metagenomic assemblies. (C) Percentage of the metatranscriptomics reads mapped to contigs with phylum, genus, and species taxonomic assignments using MyTaxa. (D) Percentage of the assemblies from July 2010 metagenomes (MTG1 and MTG2) detected in each metatranscriptomic library.

A **B**

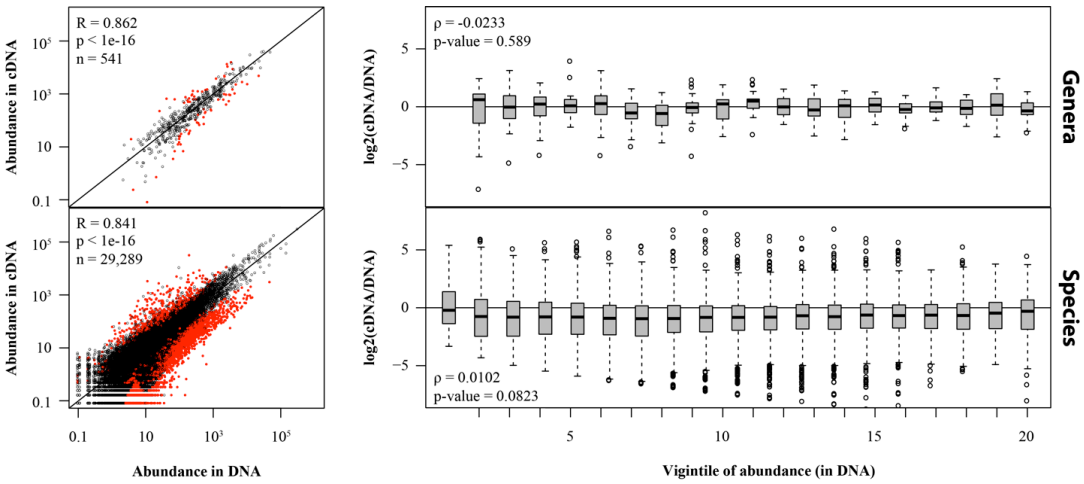
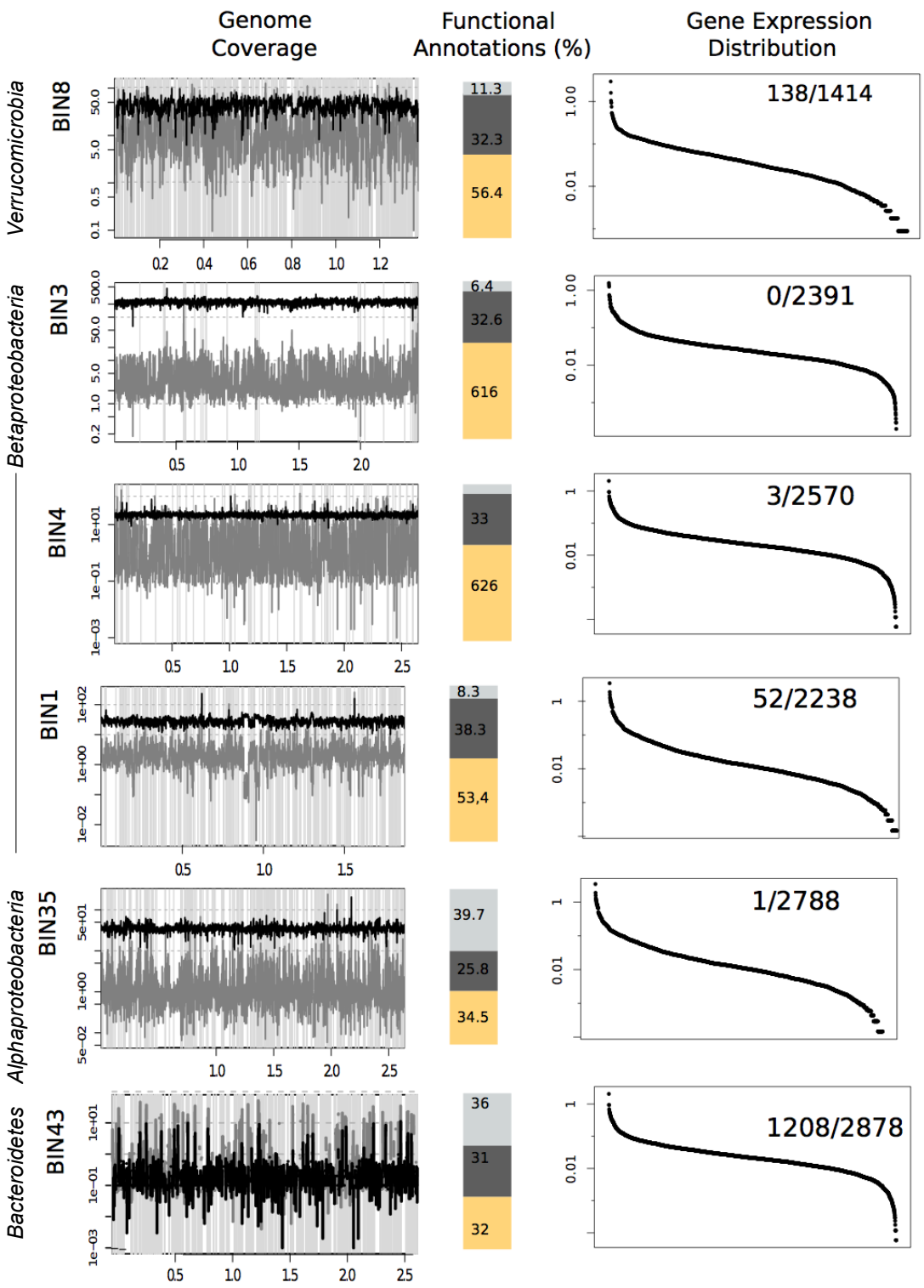


Figure B6. Relationship of cell abundance and gene transcriptional activity. Correlations were tested at two different levels: identified genera, which recruited ~16% of the total cDNA reads, and contig clusters (binned by tetra-nucleotide frequency, see Materials and Methods), which roughly represented species -including uncharacterized species- and recruited >50% of the cDNA reads. **(A)** Correlation of cDNA and DNA relative abundance for identified genera (upper panel, similar to Figure 3) or for contigs clusters (lower panel). Relative abundance refers to the total reads with annotations, i.e., reads that map to contigs that were taxonomically classified at the genus level (upper panel) or to any contig (lower panel). Red circles represent genera or species significantly overrepresented in the cDNA or DNA datasets (adjusted p-value < 0.05). **(B)** Distributions of relative expression ratios ($\log_2$(cDNA/DNA)) per vigintiles (20-quantiles) of DNA abundance for identified genera and species. The monotonicity of the correlation of DNA and cDNA abundance was tested by estimating the Spearman correlation (ρ) between $\log_2$(cDNA/DNA) values and DNA abundance. For each vigintile the distribution of relative expression ratios is displayed as the median (solid black line), the 1st to 4th inter-quartile range (grey box), 1.58 inter-quartile range (dashed lines), and outliers (circles). Note that for all ranges of abundance, the distributions were similar and centered around zero, suggesting that -on average- organisms contribute to community transcriptome proportional to their DNA abundance and there is no significant difference between low and high abundance populations (monotonicity). These results were reproducible with different quantile partitions (deciles, vigintiles, centiles, or permiles).


Genome Coverage
Functional Annotations (%)
Gene Expression Distribution
Verrucomicrobia
BIN8
11.3
32.3
56.4
138/1414
Betaproteobacteria
BIN3
6.4
32.6
616
0/2391
BIN4
33
626
3/2570
BIN1
8.3
38.3
53,4
52/2238
Alphaproteobacteria
BIN35
39.7
25.8
34.5
1/2788
Bacteroidetes
BIN43
36
31
32
1208/2878

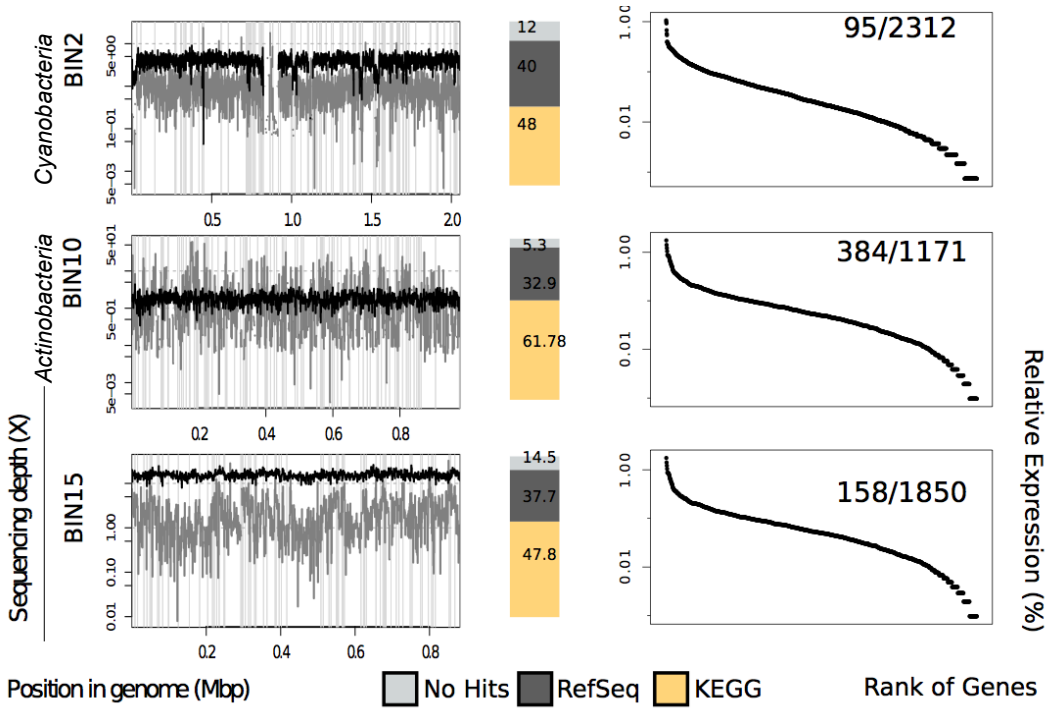


Figure B7. Recovered population genomes. The left panels display the read coverage across the recovered population genome in the July 2010 metagenome dataset (MTG1) based on the number of reads mapping with identity over the total read length ≥ 95% (black) or between 70% and 95% (light grey). Middle panels indicate the fraction of genes in the genome annotated using KEGG, RefSeq, or lacking annotation (see figure key). Right panels show the rank-abundance distribution of gene transcripts for each population genome. The numbers denote the number of genes in the genome not detected in the metatranscriptomic datasets out of the total predicted genes. Note that only the nine genomes shown met the quality criteria to be considered for further analysis (see Table S7).

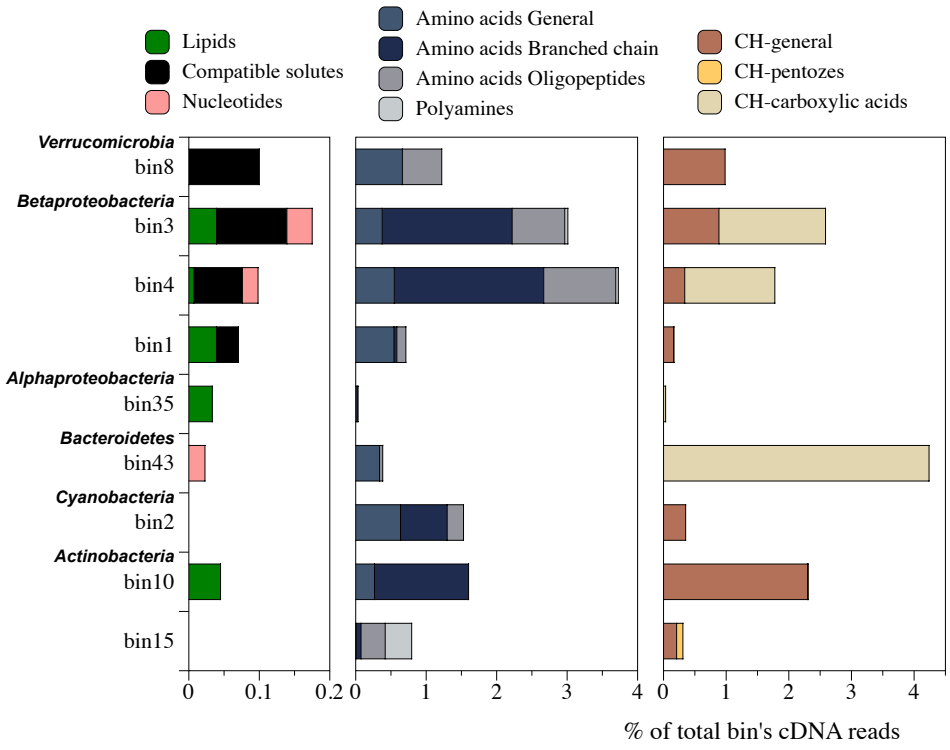


Figure B8. Expression levels of dissolved carbon transporters in recovered population genomes. The relative expression levels were calculated separately for each bin based on the sum of cDNA reads assigned to all genes constituting a transporter according to the COG gene annotation over the total reads assigned to the given population genome (bin).

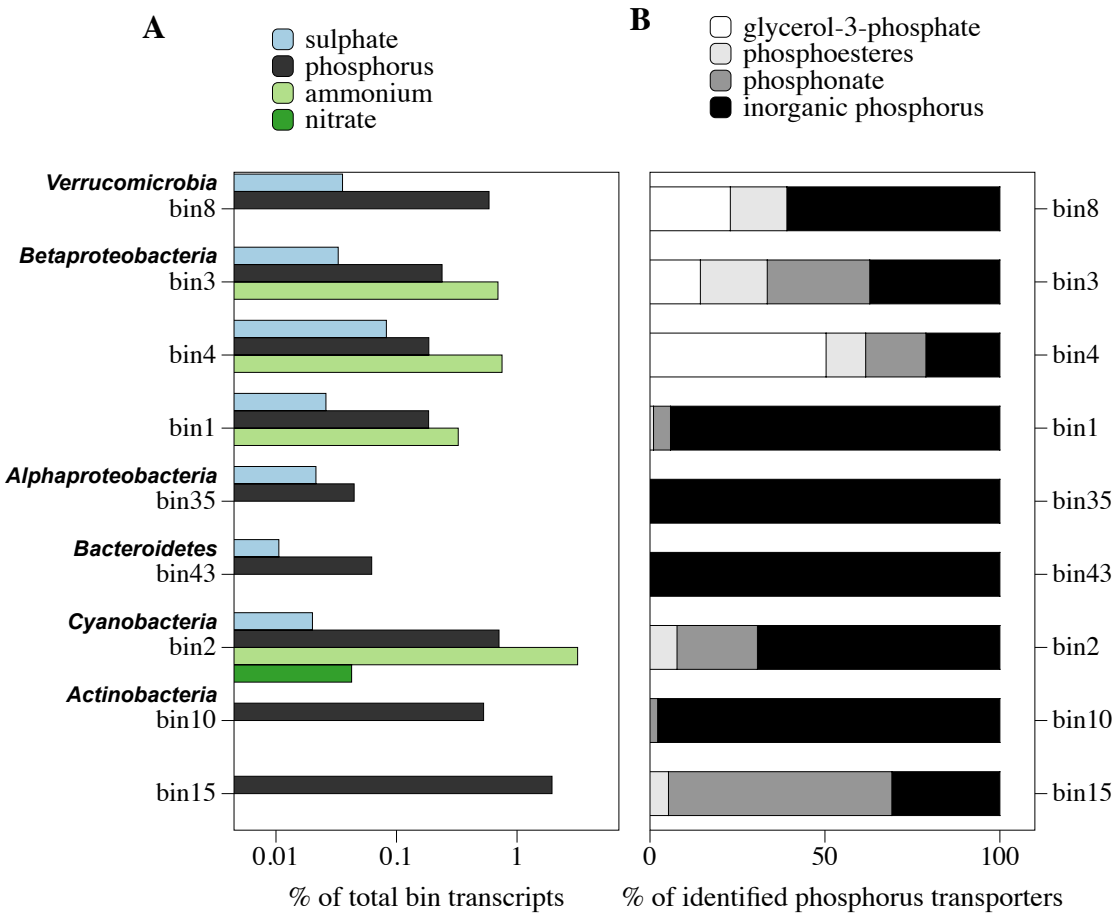


Figure B9. Expression levels of essential nutrient transporters in recovered population genomes. (A) Fraction of the total cDNA reads assigned to each population genome (bin) that encoded nutrient transporters. Results are based on keyword searches in the list of gene annotations provided by the Blast2GO pipeline and GO terms, which identified transporters as follows: **Sulphate:** suplhate permease, sulphate abc transporter substrate binding protein, sulphate abc transporter inner membrane subunit, sulphate transporter AND GO=P:sulfate transport; F:sulfate transmembrane transporter activity. **Ammonium:** ammonium transporter AND GO= F:ammonium transmembrane transporter activity, P:ammonium transmembrane transport. **Nitrate:** P:nitrate transport; F:nitrate transmembrane transporter activity. **Phosphorus**: includes all categories depicted in panel B. (**B**) Relative expressions of identified phosphorus transporters. **Glycerol-3-phosphate:** Glycerol-3-phosphate permease, Glycerol-3-phosphate permease transport system/membrane protein AND GO= P:transmembrane transport, F=trans porter activity. **Phosphoesteres:** alkaline phosphatase.

Phosphonate: phosphonate abc-transporter periplasmic/inner membrane protein, permease GO= F:organic phosphonate transmembrane transporter activity; C:integral to plasma membrane; P:organic phosphonate transport. **Inorganic phosphorus:** phosphate transporter, phosphate abc transporter permease/substrate binding protein AND GO P:phosphate ion transmembrane transport, F:inorganic phosphate transmembrane transporter activity, F:transporter activity.

SECTION A.4: REFERENCES

1. Suzuki, M. T. *et al.* Phylogenetic screening of ribosomal RNA gene-containing clones in Bacterial Artificial Chromosome (BAC) libraries from different depths in Monterey Bay. *Microb. Ecol.* **48,** 473–488 (2004).
2. Poretsky, R. S., Sun, S., Mou, X. & Moran, M. A. Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environ. Microbiol.* **12,** 616–627 (2010).
3. Oh, S. *et al.* Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Appl. Environ. Microbiol.* **77,** 6000–6011 (2011).
4. Frias-Lopez, J. *et al.* Microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 3805–3810 (2008).
5. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
6. Urban, S. G. and D. *ecodist: Dissimilarity-based functions for ecological analysis*. (2013).
7. Gilbert, J. A. *et al.* Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PloS One* **3,** e3042 (2008).
8. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet. TIG* **16,** 276–277 (2000).
9. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinforma. Oxf. Engl.* **28,** 3150–3152 (2012).
10. Williamson, S. J. *et al.* The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PloS One* **3,** e1456 (2008).
11. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinforma. Oxf. Engl.* **27,** 764–770 (2011).
12. Oberg, A. L., Bot, B. M., Grill, D. E., Poland, G. A. & Therneau, T. M. Technical and biological variance structure in mRNA-Seq data: life in the real world. *BMC Genomics* **13,** 304 (2012).

13. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* **26,** 139–140 (2010).
14. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* (2013). doi:10.1038/nbt.2579
15. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20,** 53–65 (1987).
16. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
17. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7,** e1002195 (2011).
18. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinforma. Oxf. Engl.* **22,** 2688–2690 (2006).
19. Kelley, D. R., Schatz, M. C. & Salzberg, S. L. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.* **11,** R116 (2010).
20. Melsted, P. & Pritchard, J. K. Efficient counting of k -mers in DNA sequences using a bloom filter. *BMC Bioinformatics* **12,** 333 (2011).
21. Hurwitz, B. L., Deng, L., Poulos, B. T. & Sullivan, M. B. Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ. Microbiol.* **15,** 1428–1440 (2013).
22. Fang, Z., Martin, J. & Wang, Z. Statistical methods for identifying differentially expressed genes in RNA-Seq experiments. *Cell Biosci.* **2,** 26 (2012).
23. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
24. Giannoukos, G. *et al.* Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol.* **13,** R23 (2012).

APPENDIX B

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

Table B1. Nutrient concentrations in three sampling sites.

Environmental parameters	kal1-nov	kal2-nov	kal3-nov	kal1-feb	kal2-feb	kal3-feb	kal1-may	kal2-may	kal3-may
Temperature (°C)	15	15	13.9	11	10	10.5	16	17	20
pH	7	6.5	7	6.5	6.5	6.5	6.5	6.5	6.5
Salinity (ppt)	0	0	15.89	0	0	1.57	0	0	15.89
day	21	21	21	1	1	1	48	48	48
chla(ug/l)	0.325	0.006	0.094	2.073	0.978	0.862	1.649	0.94	0.563
chlb(ug/l)	0	0	0	1.995	0.487	0.11	0.248	0.025	0.084
chlc(ug/l)	0	0	0	2.659	0.91	1.655	0.718	1.381	1.056
caroten(ug/l)	0	0	0	0	0	0	0.414	0.104	0.093
chla/(chla+pheo)	0.67	1	0.222	1	0.976	1	0.911	0.758	1
Distance from the exit of the ditch (m)	372	68,669	79,826	372	68,669	79,826	372	68,669	79826
WGS dataset statistics	kal1-nov	kal2-nov	kal3-nov	kal1-feb	kal2-feb	kal3-feb	kal1-may	kal2-may	kal3-may
#paired Reads(trimmed)	16,698,388	17,143,598	14,542,639	6,393,851	5,699,532	13,316,376	14,608,107	5,404,172	8,200,432
# predicted genes	217,938	222,836	211,639	94,388	63,909	120,245	160,935	51,688	79,471
# contigs	117,717	131,175	129,148	53,320	39,876	73,957	100,487	31,858	44,625
Coverage	0.77	0.8	0.69	0.51	0.36	0.5	0.71	0.7	0.79
N50	1,224	1,013	946	949	813	883	904	938	1,161
#reads with hits on genes	14,661,179	17,605,133	11,245,907	3,519,183	1,767,590	4,891,394	13,446,558	4,888,615	8,741,997
% reads Mapping on genes	43.89	51.34	38.66	27.52	15.5	18.36	46.02	45.23	53.3
% reads classified at the phylum level with Mytaxa	84.93	94.77	94.97	60.84	76.7	82.39	97.32	97.59	97.93
% reads classified in subsystems	22.53	31.05	40.14	20.96	47.18	43.13	27.23	29.13	38.1
% reads classified on genus level with Mytaxa	44.5	40.86	37.97	52.06	40.72	38.5	62.08	78.43	61.41
% reads classified on species level with Mytaxa	24.74	18.52	17.42	35.08	21.96	20.3	23.64	21.64	29.32

Table B2. Environmental parameters and dataset statistics.

		BOD5	COD	SS	T-N	NH4-N	NO3-N	T-P
NOVEMBER	20/11/2012		24.7	11	9.4	0.65	7.9	0.92
	19/11/2012	8.5	37.3	6.4	9.5	0.7	7.9	0.91
	16/11/2012		27.5	9	7.8	0.4	6.6	0.93
	15/11/2012		19.1	8.4	10.1	1.53	7.6	0.9
	14/11/2012		20.6	11	9.4	1.9	6.6	0.93
	13/11/2012	5.6	19.8	8.8	9.1	0.94	7.3	1.04
	12/11/12		7.5	3.8	8.4	0.12	7.5	0.9
	Average	**7.1**	**22.4**	**8.3**	**9.1**	**0.9**	**7.3**	**0.9**
	St. deviation	**2.1**	**9.1**	**2.6**	**0.8**	**0.6**	**0.6**	**0.0**
		BOD5	COD	SS	T-N	NH4-N	NO3-N	T-P
FEBRUARY	26/2/2013		14.4	6	3.7	0.22	3.1	0.81
	25/2/2013	2.9	11.8	4	7.1	0.04	6.4	0.8
	22/2/2013		11.2	3	9.2	0.09	8.2	0.85
	20/2/2013		15.6	4.2	9.7	0.08	8.8	0.86
	19/2/2013		14.1	4.2	8.7	0.07	7.8	0.83
	18/2/2013	5.1	18.4	3.8	9.3	0.08	8.3	0.82
	Average	**4.0**	**14.3**	**4.2**	**8.0**	**0.1**	**7.1**	**0.8**
	St. deviation	**1.6**	**2.6**	**1.0**	**2.3**	**0.1**	**2.1**	**0.0**
		BOD5	COD	SS	T-N	NH4-N	NO3-N	T-P
MAY	27/5/2013	3.8	14	3.6	5.3	0.07	4.7	0.69
	24/5/2013		17.5	2.8	4.8	0.52	3.8	0.55
	23/5/2013		22.9	4.6	5.6	0.31	4.8	0.65
	22/5/2013		23.8	5.2	4.8	0.54	4.2	0.91
	21/5/2013		24.4	3.6	5.9	0.11	5.2	0.95
	20/5/2013	6.5	23.3	5.8	5.6	0.11	5	0.85
	Average	**5.2**	**21.0**	**4.3**	**5.3**	**0.3**	**4.6**	**0.8**
	St. deviation	**1.9**	**4.2**	**1.1**	**0.5**	**0.2**	**0.5**	**0.2**

APPENDIX C

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

Table C1: Sequence and quality statistics for natural populations genomes

BIN ID	ORIGIN	Bin size (bp)	Taxonomy	Completeness (%)	Contamination (%)	CCI
AD-A	Nitrate Bioreactor	2918349	k__Bacteria	74.47	0.43	0.9527627
AD-J	Nitrate Bioreactor	4165136	f__Rhodobacteraceae	85.63	0.95	0.9778709
SCN_Bord	Cyanate Biorector	4799183	p__Proteobacteria	65.39	4.13	0.9720349
SCN_Micr	Cyanate Biorector	2638424	o__Actinomycetales	94.71	1.25	0.977
SCN_Pela	Cyanate Biorector	3955844	o__Rhizobiales	96.04	0.69	0.983
SCN-act	Cyanate Biorector	3122347	o__Actinomycetales	98.99	0	0.98
SCN-thio2	Cyanate Biorector	2628916	c__Betaproteobacteria	91.13	2.9	0.917
ANA-CKuenial	Anammox	3764664	k__Bacteria	94.99	3.91	0.961
BR-DTU010	Thermophillic reactor	2067643	p__Firmicutes	91.96	1.98	0.961
BR-DTU013	Thermophillic reactor	2130056	o__Clostridiales	89.34	0.7	0.977
BR-DTU014	Thermophillic reactor	2266387	p__Firmicutes	94.12	0.49	0.963
BR-DTU015	Thermophillic reactor	1735143	p__Firmicutes	90.26	0.99	0.97
BR-DTU022	Thermophillic reactor	1983120	p__Firmicutes	89.62	0.85	0.955
BR-DTU023	Thermophillic reactor	1395177	o__Clostridiales	80.38	1.42	0.988
BR-DTU024	Thermophillic reactor	2487536	c__Clostridia	95.54	2.96	0.977
BR-DTU025	Thermophillic reactor	2005558	p__Firmicutes	95.1	2.45	0.962
BR-DTU030	Thermophillic reactor	1811268	p__Firmicutes	94.89	0.42	0.962
BR-DTU052	Thermophillic reactor	2282100	c__Clostridia	97.15	1.09	0.968
BR-DTU054	Thermophillic reactor	2247352	o__Clostridiales	91.42	1.05	0.973
SE-BurM3A	Multiple amendements	9407616	k__Bacteria	80.52	0	0.97
SE-BurM3B	Multiple amendements	8810253	k__Bacteria	81.72	3.45	0.922
SE-Kte	Multiple amendements	6350800	k__Bacteria	98.68	4.36	0.975
SE-Lei2	Multiple amendements	3786980	o__Actinomycetales	97.9	1.01	0.978
SE-RhoMT2	Multiple amendements	4010263	f__Xanthomonadaceae	81.88	0.82	0.956
SE-Spo1	Multiple amendements	3484120	p__Firmicutes	98.45	0	0.982
SE-Spo2	Multiple amendements	3356144	p__Firmicutes	98.45	0.78	0.983
SH-Myc1	Hospital	6523049	o__Actinomycetales	97.73	1.14	0.975
Bank.001	Bankia	2981788	c__Gammaproteobacteria	81.96	1.77	0.897
mOL-004	Olavious	2354331	c__Gammaproteobacteria	82.6	0	0.851
mOL-006	Olavious	5832280	c__Deltaproteobacteria	99.35	3.55	0.919
OA-003	Olavious	3850915	k__Bacteria	20.8	0.16	0.6313713
TI-CoxGS3	Tick	1566493	c__Gammaproteobacteria	90.7	0.58	0.936
HG-MGS108	Human Gut	1633002	o__Clostridiales	98.66	0.67	0.974
HG-MGS119	Human Gut	1924354	f__Bifidobacteriaceae	94.82	0	0.64
HG-MGS194	Human Gut	2591406	f__Lachnospiraceae	99.38	2.42	0.954
HG-MGS221	Human Gut	3093482	o__Clostridiales	97.83	0.81	0.977
HG-MGS222	Human Gut	3749602	k__Bacteria	65	0	0.9786057
HG-MGS267	Human Gut	1723311	k__Bacteria	93.41	0	0.892
HG-MGS314	Human Gut	1508874	c__Clostridia	93.95	0.81	0.937
HG-MGS368	Human Gut	1602574	p__Actinobacteria	96.59	0.12	0.975
HG-MGS435	Human Gut	2894688	p__Bacteroidetes	93.49	1.83	0.863
HG-MGS510	Human Gut	2767136	f__Lachnospiraceae	97.73	1.55	0.914
HGS-MG257	Human Gut	2852468	o__Clostridiales	95.57	0.21	0.973
HGS-MG317	Human Gut	2328028	f__Lachnospiraceae	94.44	0.29	0.966
HGS-MG424	Human Gut	3017849	o__Clostridiales	98.16	1.27	0.979
HGS-MG522	Human Gut	1355207	c__Clostridia	95.56	1.61	0.954
HGS-MG534	Human Gut	2677546	f__Lachnospiraceae	97.41	2.37	0.954
HGS-MG59	Human Gut	5406274	k__Bacteria	73.64	0	0.9766996
HGS-MG633	Human Gut	3636452	o__Bacteroidales	96.08	1.86	0.899

GW-1618478	Groundwater Sed.	1118955	k__Bacteria	62.13	1.12	0.9668612
GW-1618482	Groundwater Sed.	1124729	k__Bacteria	61.27	2.31	0.9692204
GW-1618485	Groundwater Sed.	1019220	k__Bacteria	51.53	2.25	0.9586545
GW-1618488	Groundwater Sed.	962636	k__Bacteria	73.82	3.45	0.9650244
GW-1618500	Groundwater Sed.	1196777	k__Bacteria	52.64	0	0.97172
GW-1618511	Groundwater Sed.	1037017	k__Bacteria	49.26	0	0.959365
GW-1618519	Groundwater Sed.	999614	k__Bacteria	60.06	2.42	0.9665398
GW-1618520	Groundwater Sed.	1076430	k__Bacteria	50.06	0	0.9681683
GW-1618539	Groundwater Sed.	977514	k__Bacteria	64.76	1.98	0.9577352
GW-1618547	Groundwater Sed.	1038907	k__Bacteria	56.84	0	0.979684
GW-1618550	Groundwater Sed.	1043864	k__Bacteria	58.58	0	0.9666845
GW-1618636	Groundwater Sed.	1387850	k__Bacteria	69.8	0	0.9665941
GW-1618637	Groundwater Sed.	1388361	k__Bacteria	69.8	0	0.9634896
GW-1618638	Groundwater Sed.	1350364	k__Bacteria	68.81	0	0.950707
GW-1618639	Groundwater Sed.	1382032	k__Bacteria	68.81	0	0.9638813
GW-1618640	Groundwater Sed.	1387108	k__Bacteria	69.8	0	0.9676408
GW-1618641	Groundwater Sed.	1379854	k__Bacteria	68.81	0	0.9602507
GW-1618696	Groundwater Sed.	1151890	k__Bacteria	89.5	0	0.9542172
GW-1618707	Groundwater Sed.	1375995	k__Bacteria	75.52	3.37	0.9334468
GW-1618715	Groundwater Sed.	1272956	k__Bacteria	86.52	0	0.9597196
GW-1618716	Groundwater Sed.	1167397	k__Bacteria	63.37	2.81	0.9563815
GW-1618722	Groundwater Sed.	1222490	k__Bacteria	68.26	1.12	0.9728711
GW-1618726	Groundwater Sed.	1137406	k__Bacteria	86.91	3.45	0.9694675
GW-1618780	Groundwater Sed.	1033177	k__Bacteria	74.01	1.85	0.9726454
GW-1618811	Groundwater Sed.	1036782	k__Bacteria	58.91	0	0.9616185
GW-1618816	Groundwater Sed.	1222335	k__Bacteria	71.78	2.97	0.9602204
GW-1618820	Groundwater Sed.	1244279	k__Bacteria	62.93	0	0.9527753
GW-1618827	Groundwater Sed.	1301823	k__Bacteria	69.8	0	0.9672505
GW-1618832	Groundwater Sed.	1288159	k__Bacteria	64.81	0.19	0.9671434
GW-1618838	Groundwater Sed.	1074186	k__Bacteria	65.74	0.99	0.9641256
GW-1618839	Groundwater Sed.	1330629	k__Bacteria	67.56	0.93	0.9620087
GW-1618850	Groundwater Sed.	1259749	k__Bacteria	67.43	0	0.9646073
GW-1618894	Groundwater Sed.	1047167	k__Bacteria	71.29	0	0.9660784
GW-1618922	Groundwater Sed.	1222923	k__Bacteria	71.78	2.97	0.9733667
GW-1618929	Groundwater Sed.	1063161	k__Bacteria	70.3	0	0.966707
GW-1618930	Groundwater Sed.	1245242	k__Bacteria	63.46	1.83	0.9581564
GW-1618944	Groundwater Sed.	1077788	k__Bacteria	63.83	1.17	1.880884
GW-1618973	Groundwater Sed.	1135682	k__Bacteria	72.66	0	0.9722541
GW-1618975	Groundwater Sed.	1115935	k__Bacteria	55.14	4.27	0.9568899
GW-1618985	Groundwater Sed.	1066372	k__Bacteria	69.8	0	0.9590864
GW-1618994	Groundwater Sed.	1055601	k__Bacteria	70.79	0.99	0.9605678
GW-1618995	Groundwater Sed.	1164937	k__Bacteria	52.48	0	0.9730147
GW-1618996	Groundwater Sed.	1068223	k__Bacteria	70.18	1.83	0.949061
GW-1619001	Groundwater Sed.	1061317	k__Bacteria	70.79	0.99	0.9537588
GW-1619040	Groundwater Sed.	1093919	k__Bacteria	77.53	1.12	0.9637418
GW-1619044	Groundwater Sed.	1249622	k__Bacteria	72.28	0	0.9466368
GW-1619051	Groundwater Sed.	1148067	k__Bacteria	79.78	1.12	0.958776
GW-1619054	Groundwater Sed.	1259947	k__Bacteria	75.49	1.12	0.9693908
GW-1619055	Groundwater Sed.	993453	k__Bacteria	61.21	0	0.9547306
GW-1619056	Groundwater Sed.	1082319	k__Bacteria	75.46	0.93	0.969596
GW-1619061	Groundwater Sed.	1433016	k__Bacteria	80.9	2.81	0.974075

ES-DG-56	Estuary Sediments	1238975	k__Bacteria	51.23	0.93	0.9672901
ES-DG-58	Estuary Sediments	1746620	k__Bacteria	47.66	1.14	0.9781884
ES-DG-60	Estuary Sediments	1357456	c__Deltaproteobacteria	61.93	0.33	0.9718592
ES-DG-78	Estuary Sediments	2090507	k__Bacteria	75.6	1.19	0.9842764
ES-SM1-40	Estuary Sediments	2559446	k__Bacteria	84.62	1.1	0.974
ES-SM1-77	Estuary Sediments	1269144	k__Bacteria	27.12	0	0.9116281
ES-SM23-28-1	Estuary Sediments	1458299	k__Bacteria	70.96	0	0.9356264
ES-SM23-33	Estuary Sediments	3426709	k__Bacteria	75	1.14	0.9765377
ES-SM23-39	Estuary Sediments	1126604	k__Bacteria	82.77	0.68	0.957
ES-SM23-42	Estuary Sediments	2532789	k__Bacteria	93.41	1.1	0.967
ES-SM23-65	Estuary Sediments	1825793	k__Bacteria	48.49	2.22	0.9685138
FBin-1	Soils	4341484	k__Bacteria	93.01	2.74	0.945
FBin-27	Soils	8218361	c__Deltaproteobacteria	88.55	4.52	0.986
FBin-3	Soils	5822639	k__Bacteria	95.51	1.71	0.977
FBin-4	Soils	2468837	k__Bacteria	70.11	0.84	0.9735151
FBin-6	Soils	3684989	k__Bacteria	93.4	1.03	0.97
FBin-7	Soils	5910590	k__Bacteria	91.78	0.85	0.971
FBin-8	Soils	3268372	p__Actinobacteria	92.02	0.85	0.982
GW-1618343_	Groundwater Sed.	1071934	k__Bacteria	63.43	0.1	0.957546
GW-1618345_	Groundwater Sed.	1098928	k__Bacteria	73.76	0.11	0.962845
GW-1618346_	Groundwater Sed.	1123966	k__Bacteria	75	0.46	0.9663388
GW-1618348_	Groundwater Sed.	1076678	k__Bacteria	70.37	0.46	0.9679171
GW-1618350_	Groundwater Sed.	1136276	k__Bacteria	71.78	0.99	0.9717019
GW-1618351_	Groundwater Sed.	1084784	k__Bacteria	71.18	4.09	0.9582192
GW-1618363_	Groundwater Sed.	1009164	k__Bacteria	58.91	0.99	0.9618504
GW-1618368_	Groundwater Sed.	1086326	k__Bacteria	66.01	0	0.933197
GW-1618370_	Groundwater Sed.	1147980	k__Bacteria	66.01	2.53	0.944798
GW-1618372_	Groundwater Sed.	1049888	k__Bacteria	66.5	0	0.9540339
GW-1618374_	Groundwater Sed.	959638	k__Bacteria	61.55	0.5	0.9649568
GW-1618379_	Groundwater Sed.	1089432	k__Bacteria	61.5	0	0.9603724
GW-1618381_	Groundwater Sed.	1095642	k__Bacteria	62.78	0	0.9676332
GW-1618387_	Groundwater Sed.	1090004	k__Bacteria	61.3	0	0.9748011
GW-1618393_	Groundwater Sed.	1035077	k__Bacteria	59.81	0.99	0.9431725
GW-1618396_	Groundwater Sed.	1027882	k__Bacteria	60.32	0	0.9652255
GW-1618398_	Groundwater Sed.	1101807	k__Bacteria	60.8	0	0.9578461
GW-1618411_	Groundwater Sed.	1046499	k__Bacteria	69.41	0	0.9538851
GW-1618415_	Groundwater Sed.	1099135	k__Bacteria	66.36	0	0.9699841
GW-1618417_	Groundwater Sed.	1103893	k__Bacteria	61.88	2.97	0.974556
GW-1618423_	Groundwater Sed.	1272388	k__Bacteria	69.44	0	0.9591589
GW-1618424_	Groundwater Sed.	1035471	k__Bacteria	59.72	1.94	0.9494751
GW-1618426_	Groundwater Sed.	1080324	k__Bacteria	65.84	2.09	0.9608738
GW-1618430_	Groundwater Sed.	1024051	k__Bacteria	55.09	1.39	0.958625
GW-1618431_	Groundwater Sed.	1271115	k__Bacteria	70.37	0	0.9765416
GW-1618435_	Groundwater Sed.	1460774	k__Bacteria	64.52	0	0.950788
GW-1618436_	Groundwater Sed.	1424122	k__Bacteria	62.38	2.97	0.9658216
GW-1618441_	Groundwater Sed.	1325052	k__Bacteria	61.55	1.32	0.9622341
GW-1618442_	Groundwater Sed.	1241278	k__Bacteria	65.02	0.36	0.9756427
GW-1618443_	Groundwater Sed.	1650629	k__Bacteria	71.95	0	0.9566388
GW-1618447_	Groundwater Sed.	1215815	k__Bacteria	53.03	2.97	0.9598567
GW-1618449_	Groundwater Sed.	1416294	k__Bacteria	67.88	0.25	0.9710019
GW-1618474_	Groundwater Sed.	1023535	k__Bacteria	59.14	3.7	0.9541176

MAG-121015-	Marine (BS)	1209326	k__Bacteria	80.92	2.83	0.956
MAG-121015-	Marine (BS)	1495089	c__Gammaproteobacteria	81.61	1.44	0.88
MAG-121022-	Marine (BS)	1495089	c__Gammaproteobacteria	81.61	1.44	0.895
MAG-121128-	Marine (BS)	1159044	k__Bacteria	77.56	0.56	0.852133
MAG-121128-	Marine (BS)	1055524	o__Rickettsiales	85.74	2.88	0.892
MAG-121128-	Marine (BS)	1469346	c__Gammaproteobacteria	81.75	1.98	0.881
MAG-121220-	Marine (BS)	1915951	s__algicola	93.58	1.08	0.966
MAG-121220-	Marine (BS)	1270387	k__Archaea	99.51	0.97	0.977
MAG-121220-	Marine (BS)	2112289	f__Rhodobacteraceae	70.84	2.14	0.9587121
MAG-121220-	Marine (BS)	979010	o__Actinomycetales	67.27	1.63	0.954842
MAG-121220-	Marine (BS)	1205145	k__Bacteria	66.32	1.74	0.9729522
MAG-121220-	Marine (BS)	1907425	c__Gammaproteobacteria	82.89	1.07	0.972
MAG-121220-	Marine (BS)	1480518	k__Bacteria	87.13	2.94	0.961
MAG-121220-	Marine (BS)	1723929	k__Bacteria	95.43	0	0.979
ALL-101	Marine (GOM)	1870925	k__Bacteria	42.01	0	0.9036972
ALL-102	Marine (GOM)	1267190	k__Bacteria	30.96	1.18	0.8488788
ALL-104	Marine (GOM)	958045	k__Bacteria	18.18	0.86	0.9342947
ALL-105	Marine (GOM)	1136755	o__Actinomycetales	63.25	1.27	0.9237712
ALL-108	Marine (GOM)	1952414	k__Bacteria	20.38	4.21	0.656132
ALL-121	Marine (GOM)	1239759	c__Betaproteobacteria	51.95	2.26	0.9427076
ALL-125	Marine (GOM)	1287494	k__Bacteria	48.02	3.62	0.924746
ALL-160	Marine (GOM)	1902438	k__Bacteria	25.16	2.87	0.6635667
ALL-165	Marine (GOM)	1228406	p__Bacteroidetes	62.26	1.1	0.9348391
ALL-179	Marine (GOM)	1558705	k__Bacteria	36.13	0	0.9650531
ALL-181	Marine (GOM)	1038293	p__Proteobacteria	24.26	2.54	0.635411
ALL-189	Marine (GOM)	1523731	k__Archaea	23.83	4.98	0.6334798
ALL-201	Marine (GOM)	1073500	k__Bacteria	48.75	1.72	0.9458904
ALL-225	Marine (GOM)	1821242	p__Proteobacteria	96.34	1.07	0.967
ALL-236	Marine (GOM)	1414546	k__Bacteria	25.09	2.49	0.639993
ALL-238	Marine (GOM)	2639280	p__Proteobacteria	94.14	1.3	0.972
ALL-287	Marine (GOM)	1142540	k__Archaea	22.03	3.89	0.6726823
ALL-316	Marine (GOM)	1143072	k__Bacteria	52.54	3.45	0.9672084
ALL-317	Marine (GOM)	1363030	k__Bacteria	32.88	3.45	0.9799285
ALL-34	Marine (GOM)	1158413	k__Bacteria	19.75	0.63	0.6342019
ALL-433	Marine (GOM)	1084422	g__Vibrio	18.75	1.22	0.9626846
ALL-50	Marine (GOM)	986616	k__Bacteria	37.93	3.45	0.9307671
ALL-66	Marine (GOM)	1023603	k__Bacteria	31.18	3.45	0.6345866
ALL-71	Marine (GOM)	1008002	k__Bacteria	57.64	1.72	0.9593409
OMZ-003	Marine (OMZ)	1263853	k__Bacteria	78.41	1.2	0.9697776
OMZ-007	Marine (OMZ)	2405161	k__Bacteria	85.51	4.4	0.943
OMZ-008	Marine (OMZ)	1827440	k__Bacteria	81.87	0.65	0.978
OMZ-028	Marine (OMZ)	1624115	p__Proteobacteria	80.89	4.03	0.965
OMZ-004	Marine (OMZ)	1896701	k__Bacteria	62.98	2.71	0.9751824
OMZ-001	Marine (OMZ)	1893142	p__Proteobacteria	91.16	3	0.959
CB-BA1	Marine Sediments	1931714	k__Archaea	91.59	2.8	0.978
CB-BA2	Marine Sediments	1455689	k__Archaea	93.77	3.74	0.974
ES-DG-20	Estuary Sediments	2821569	k__Bacteria	58.52	4.24	0.9687612
ES-DG-22	Estuary Sediments	1135001	k__Bacteria	46.32	0.99	0.9589428
ES-DG-23	Estuary Sediments	2020164	k__Bacteria	82.95	2.34	0.968
ES-DG-24	Estuary Sediments	2323720	k__Bacteria	63.74	0	0.9443844
ES-DG-27	Estuary Sediments	1101969	k__Bacteria	56.35	0	0.9626765

LL-92	Freshwater	960935	k__Bacteria	49.76	0	0.9341563
LL-93	Freshwater	991959	o__Actinomycetales	61.95	0.36	0.9663694
LL-96	Freshwater	2455305	o__Burkholderiales	96.54	1.17	0.971
MAG-120322-	Marine (BS)	2980252	k__Bacteria	97.3	4.05	0.871
MAG-120322-	Marine (BS)	1262816	o__Actinomycetales	84.09	0	0.97
MAG-120322-	Marine (BS)	1743356	k__Bacteria	97.85	0	0.954
MAG-120322-	Marine (BS)	1753605	k__Bacteria	88.62	2.99	0.933
MAG-120322-	Marine (BS)	1997685	c__Gammaproteobacteria	87.88	1.84	0.956
MAG-120507-	Marine (BS)	1482147	c__Gammaproteobacteria	69.27	1.85	0.8824516
MAG-120507-	Marine (BS)	1601779	k__Bacteria	85	1.35	0.965
MAG-120507-	Marine (BS)	1181071	k__Bacteria	65.05	2.15	0.963612
MAG-120507-	Marine (BS)	2338460	c__Gammaproteobacteria	82.55	1.57	0.968
MAG-120531-	Marine (BS)	2086531	s__algicola	93.65	0.66	0.843
MAG-120531-	Marine (BS)	1121591	o__Actinomycetales	87.99	3.12	0.952
MAG-120531-	Marine (BS)	2244757	c__Gammaproteobacteria	82.44	2.67	0.935
MAG-120619-	Marine (BS)	2648901	c__Gammaproteobacteria	89.1	2.28	0.948
MAG-120619-	Marine (BS)	2408986	f__Flavobacteriaceae	91.77	1.24	0.962
MAG-120619-	Marine (BS)	1498937	s__algicola	77.99	3.53	0.8801496
MAG-120619-	Marine (BS)	2527476	c__Gammaproteobacteria	97.78	0	0.958
MAG-120619-	Marine (BS)	1214420	o__Actinomycetales	83.48	2.02	0.937
MAG-120813-	Marine (BS)	1156631	o__Actinomycetales	73.04	4.57	0.8991672
MAG-120813-	Marine (BS)	1264266	c__Gammaproteobacteria	67.65	0.95	0.9542889
MAG-120813-	Marine (BS)	2799617	p__Bacteroidetes	91.94	0.74	0.975
MAG-120820-	Marine (BS)	1127841	o__Actinomycetales	76.9	2.11	0.954174
MAG-120820-	Marine (BS)	1151850	o__Rickettsiales	95.32	4.27	0.958
MAG-120820-	Marine (BS)	1406795	k__Bacteria	74.94	1.35	0.8939019
MAG-120820-	Marine (BS)	1455539	c__Gammaproteobacteria	71.31	0.56	0.9048203
MAG-120820-	Marine (BS)	1091054	o__Actinomycetales	58.87	4.82	0.9007372
MAG-120823-	Marine (BS)	1672064	k__Bacteria	81.71	4.32	0.89
MAG-120823-	Marine (BS)	1352430	k__Bacteria	76.96	0.85	0.9561122
MAG-120823-	Marine (BS)	1001092	o__Actinomycetales	69.94	3.46	0.9655038
MAG-120823-	Marine (BS)	1451966	c__Gammaproteobacteria	69.85	2.04	0.8951506
MAG-120828-	Marine (BS)	1029940	c__Gammaproteobacteria	72.03	1.44	0.9734519
MAG-120910-	Marine (BS)	1746953	k__Bacteria	98.39	0.83	0.962
MAG-120910-	Marine (BS)	2763624	f__Rhodobacteraceae	91.15	3.03	0.899
MAG-120910-	Marine (BS)	1688323	k__Bacteria	90.6	3.85	0.98
MAG-120910-	Marine (BS)	1086213	c__Betaproteobacteria	88.7	3.3	0.966
MAG-120920-	Marine (BS)	1139720	o__Actinomycetales	58.83	3.1	0.8646698
MAG-120920-	Marine (BS)	2191795	c__Gammaproteobacteria	83.5	3.57	0.963
MAG-120920-	Marine (BS)	1450272	c__Gammaproteobacteria	70.57	3.85	0.8738532
MAG-120920-	Marine (BS)	1080997	o__Actinomycetales	75.62	2.18	0.9021689
MAG-120924-	Marine (BS)	1459716	k__Bacteria	87.06	0.73	0.97
MAG-120924-	Marine (BS)	2308797	c__Gammaproteobacteria	87.66	1.62	0.966
MAG-120924-	Marine (BS)	1727489	k__Bacteria	95.7	2.41	0.952
MAG-120924-	Marine (BS)	1485032	k__Bacteria	83.35	1.73	0.968
MAG-120924-	Marine (BS)	1314100	c__Gammaproteobacteria	68.93	0.49	0.911704
MAG-121001-	Marine (BS)	1655662	c__Gammaproteobacteria	65.07	1.63	0.9654293
MAG-121001-	Marine (BS)	1208227	k__Bacteria	76.52	3.94	0.9570927
MAG-121001-	Marine (BS)	1509054	c__Gammaproteobacteria	77.59	1.98	0.8843107
MAG-121001-	Marine (BS)	1121384	o__Actinomycetales	74.84	3.15	0.8893107
MAG-121001-	Marine (BS)	1043961	o__Actinomycetales	58.46	4.87	0.9039695

LL-169	Freshwater	1131204	o__Actinomycetales	75.93	1.17	0.9459182
LL-17	Freshwater	1790919	o__Burkholderiales	83.13	1.01	0.888
LL-191	Freshwater	2351168	p__Bacteroidetes	93.24	0.95	0.977
LL-192	Freshwater	3730914	k__Bacteria	79.86	1.16	0.9775046
LL-2	Freshwater	1289299	o__Actinomycetales	74.19	4.82	0.8763584
LL-20	Freshwater	1791539	c__Betaproteobacteria	96.59	0.71	0.968
LL-203	Freshwater	2324733	o__Burkholderiales	85.07	0.99	0.966
LL-208	Freshwater	2188017	p__Proteobacteria	39.5	1.02	0.9531017
LL-215	Freshwater	3547470	o__Burkholderiales	90.67	3.43	0.974
LL-229	Freshwater	1273514	o__Burkholderiales	46.76	0.62	0.9525449
LL-23	Freshwater	1744980	k__Bacteria	86.09	4.27	0.939
LL-24	Freshwater	1121016	k__Bacteria	48.15	1.72	0.8546561
LL-249	Freshwater	4832833	k__Bacteria	69.63	4.11	0.6339315
LL-25	Freshwater	2086386	p__Cyanobacteria	98.41	0.54	0.884
LL-250	Freshwater	4390298	k__Bacteria	86.09	0	0.971
LL-257	Freshwater	2539226	o__Burkholderiales	83.7	0.79	0.982
LL-27	Freshwater	1159499	o__Actinomycetales	79.4	0	0.9319913
LL-271	Freshwater	1190067	k__Bacteria	43.21	0.68	0.9640749
LL-278	Freshwater	2039566	k__Bacteria	76.84	0.06	0.9773351
LL-283	Freshwater	2118703	k__Bacteria	56.51	0.64	0.9786575
LL-29	Freshwater	4219552	k__Bacteria	98.57	2.2	0.884
LL-291	Freshwater	1431099	c__Gammaproteobacteria	68.5	0.97	0.8238976
LL-299	Freshwater	1473989	k__Bacteria	61.28	0.54	0.9611527
LL-309	Freshwater	1834331	k__Bacteria	44.83	1.25	0.6328277
LL-310	Freshwater	1202182	k__Bacteria	45.66	1.68	0.973979
LL-314	Freshwater	2274349	c__Gammaproteobacteria	93.31	2.12	0.94
LL-325	Freshwater	1714458	p__Bacteroidetes	44.99	2.49	0.8716874
LL-329	Freshwater	2267846	k__Bacteria	76.59	0	0.9640511
LL-344	Freshwater	1778555	k__Bacteria	39.84	2.47	0.6317773
LL-35	Freshwater	986345	k__Bacteria	36.9	3.45	0.9054995
LL-350	Freshwater	5337767	k__Bacteria	29.47	1.65	0.8792243
LL-352	Freshwater	2247386	k__Bacteria	43.33	2.94	0.7979558
LL-38	Freshwater	2482613	k__Bacteria	95.95	1.35	0.893
LL-39	Freshwater	1822490	k__Bacteria	87.18	2.28	0.888
LL-40	Freshwater	1545524	k__Bacteria	79.06	2.2	0.980305
LL-41	Freshwater	1872867	o__Burkholderiales	58.21	3.24	0.9090036
LL-42	Freshwater	2077781	k__Bacteria	79.26	1.85	0.8718127
LL-43	Freshwater	1724260	k__Bacteria	60.1	2.54	0.8151412
LL-47	Freshwater	1208016	o__Actinomycetales	87.34	1.8	0.91
LL-51	Freshwater	1785656	k__Bacteria	75.7	2.18	0.9686637
LL-63	Freshwater	1878922	p__Bacteroidetes	94.58	0.49	0.969
LL-64	Freshwater	1963358	k__Bacteria	81.26	1.23	0.96
LL-66	Freshwater	1043310	k__Bacteria	32.21	0	0.9702826
LL-69	Freshwater	943707	k__Bacteria	40.52	0	0.9120683
LL-70	Freshwater	1677608	k__Bacteria	81.96	2.99	0.96
LL-72	Freshwater	1432070	o__Actinomycetales	80.71	0.53	0.964
LL-73	Freshwater	1438616	k__Bacteria	41.52	0	0.9755102
LL-76	Freshwater	1225148	o__Actinomycetales	63.51	1.05	0.8576923
LL-81	Freshwater	1187071	o__Actinomycetales	59.36	0.53	0.8626447
LL-89	Freshwater	2116800	o__Burkholderiales	67.52	2.09	0.9641103
LL-9	Freshwater	3044491	p__Proteobacteria	98.11	0.62	0.861

HGS-MGS100	Human Gut	1121351	k__Bacteria	91.57	2.81	0.968
HGS-MGS118	Human Gut	3786236	o__Bacteroidales	99.03	0.95	0.972
HGS-MGS119	Human Gut	1064680	k__Bacteria	92.13	2.41	0.962
HGS-MGS120	Human Gut	3171669	o__Bacteroidales	96.66	2.06	0.971
HGS-MGS143	Human Gut	1561166	c__Clostridia	93.95	1.89	0.948
HGS-MGS157	Human Gut	2171780	p__Bacteroidetes	98.56	0.96	0.904
HGS-MGS162	Human Gut	3138897	o__Bacteroidales	97.64	0.42	0.981
HGS-MGS236	Human Gut	1630951	g__Streptococcus	97.16	0.22	0.962
HGS-MGS260	Human Gut	1809021	k__Bacteria	95.7	0	0.956
HGS-MGS274	Human Gut	2185031	o__Clostridiales	98.66	0.67	0.93
HGS-MGS28	Human Gut	2649188	f__Lachnospiraceae	97.46	0.19	0.975
HGS-MGS298	Human Gut	1784438	p__Actinobacteria	98.39	1.92	0.93
HGS-MGS303	Human Gut	2699999	o__Clostridiales	98.43	1.62	0.948
HGS-MGS348	Human Gut	2523991	o__Clostridiales	98.94	0.97	0.971
HGS-MGS409	Human Gut	3484794	o__Bacteroidales	97.69	2.31	0.978
HGS-MGS433	Human Gut	1346943	k__Bacteria	93.82	0.56	0.96
HGS-MGS470	Human Gut	1577465	o__Clostridiales	75.32	1.79	0.9096843
HGS-MGS471	Human Gut	2686678	o__Clostridiales	97.89	1.27	0.949
HGS-MGS52	Human Gut	2823805	o__Clostridiales	96.78	1.9	0.955
HGS-MGS557	Human Gut	1112347	o__Clostridiales	80.06	1.8	0.958
HGS-MGS700	Human Gut	2452409	c__Spirochaetia	97.2	0	0.978
HGS-MGS777	Human Gut	2251896	c__Gammaproteobacteria	98.79	0.44	0.962
HGS-MGS82	Human Gut	2065307	o__Clostridiales	93.25	0	0.926
HGS-MGS826	Human Gut	1584863	k__Bacteria	98.88	0.22	0.973
HGS-MGS831	Human Gut	2568546	p__Bacteroidetes	95.95	0.95	0.636
HGS-MGS882	Human Gut	1255910	k__Bacteria	93.26	4.49	0.979
HGS-MGS917	Human Gut	1529412	o__Clostridiales	90.14	1.74	0.962
HGS-MGS977	Human Gut	1923887	o__Rhodospirillales	80.4	1.04	0.938
LL-1	Freshwater	3629038	p__Proteobacteria	98.77	3.5	0.811
LL-100	Freshwater	1957708	p__Bacteroidetes	90.32	0.48	0.964
LL-102	Freshwater	1634247	k__Bacteria	48.51	1.72	0.9681796
LL-105	Freshwater	1399130	o__Actinomycetales	89.11	0.6	0.949
LL-106	Freshwater	1291964	o__Burkholderiales	47.29	0.93	0.9758251
LL-111	Freshwater	2474823	k__Bacteria	88.24	3.36	0.912
LL-112	Freshwater	1416019	k__Bacteria	76.33	0.95	0.9608117
LL-114	Freshwater	2187968	o__Burkholderiales	89.39	1.95	0.969
LL-12	Freshwater	1215533	o__Actinomycetales	84.66	0.53	0.93
LL-123	Freshwater	2217991	k__Bacteria	81.89	1.49	0.971
LL-124	Freshwater	1716025	k__Bacteria	41.54	0.07	0.9658884
LL-128	Freshwater	1471010	k__Bacteria	79.2	0.57	0.9759204
LL-129	Freshwater	1205586	k__Bacteria	34.64	0	0.9551471
LL-13	Freshwater	1023777	o__Actinomycetales	70.58	1.13	0.9559557
LL-130	Freshwater	1692629	c__Betaproteobacteria	95.02	0.17	0.895
LL-137	Freshwater	2881106	o__Burkholderiales	91.12	3.05	0.973
LL-14	Freshwater	1748740	k__Bacteria	92.57	2.03	0.976
LL-142	Freshwater	3078877	c__Gammaproteobacteria	93.68	3.01	0.898
LL-149	Freshwater	1296934	k__Bacteria	60.37	1.17	0.980263
LL-153	Freshwater	1237091	p__Proteobacteria	85.85	0	0.963
LL-157	Freshwater	1171141	k__Bacteria	54.08	0	0.9556157
LL-16	Freshwater	1263263	k__Bacteria	80.07	0.04	0.977
LL-160	Freshwater	1312511	k__Bacteria	52.11	1.72	0.9639982

APPENDIX D

SUPPLEMENTARY MATERIAL FOR CHAPTER 5

Table D1: List of enzymes examined for characteristc pathways

Botin Biosynthesis

bioF	adenosylmethionine---8-amino-7-oxononanoate aminotransferase [EC:2.6.1.62]
bioA, bioK	lysine---8-amino-7-oxononanoate aminotransferase [EC:2.6.1.105]
bioD	dethiobiotin synthetase [EC:6.3.3.3]
bioB	biotin synthase [EC:2.8.1.6]

Adenosylcobalamin Biosynthesis (B12)

De Novo branch of pathway

CysG/UroM	Uroporphyrinogen-III methyltransferase (EC 2.1.1.107)
CysG	Precorrin-2 oxidase (EC 1.3.1.76)
CbiK	Sirohydrochlorin cobaltochelatase CbiK (EC 4.99.1.3)
CbiL	Cobalt-precorrin-2 C20-methyltransferase (EC 2.1.1.130)
CbiH	Cobalt-precorrin-3b C17-methyltransferase
CbiG	Cobalamin biosynthesis protein CbiG
CbiD	Cobalt-precorrin-6 synthase, anaerobic
CbiJ	Cobalt-precorrin-6x reductase (EC 1.3.1.54)
CbiE	Cobalt-precorrin-6y C5-methyltransferase (EC 2.1.1.-)
CbiT	Cobalt-precorrin-6y C15-methyltransferase [decarboxylating] (EC 2.1.1.-)
CbiC	Cobalt-precorrin-8x methylmutase (EC 5.4.1.2)

Cobirinate branch

CbiA	Cobyrinic acid A,C-diamide synthase
BluB	Cobalamin biosynthesis protein BluB (EC 1.16.8.1)

BtuR	Cob(I)alamin adenosyltransferase (EC 2.5.1.17)
PduO	Cob(I)alamin adenosyltransferase PduO (EC 2.5.1.17)
CbiP	Cobyric acid synthase
CbiZ	Adenosylcobinamide amidohydrolase (EC 3.5.1.90)
CbiB	Adenosylcobinamide-phosphate synthase

Cobalamin uptake

BtuR	Cob(I)alamin adenosyltransferase (EC 2.5.1.17)

Cobinamide uptake (requires BtuR)

CobU	Adenosylcobinamide-phosphate guanylyltransferase (EC 2.7.7.62)
CobS	Cobalamin synthase

DMB+NMD uptake

CobT	Nicotinate-nucleotide--dimethylbenzimidazole phosphoribosyltransferase (EC 2.4.2.21)
CobC	Alpha-ribazole-5-phosphate phosphatase (EC 3.1.3.73)
CobS	Cobalamin synthase

Protection from Oxygen

Oxidases (O2 to H2O or H2O2) Feroxidases

(EC 1.1.3.15)	Glycolate oxidase
(EC 1.9.3.1)	Cytochrome c oxidase polypeptide (3 subunits)
(EC 1.10.3.-)	Cytochrome d ubiquinol oxidase subunit (2 subunits)
(EC 1.10.3.-)	Cytochrome O ubiquinol oxidase (2 subunits)
(EC 1.16.3.1)	Feroxidases

Catalases & Peroxidases

(EC 1.11.1.6)	peroxidase/catalase HPI
(EC 1.11.1.7)	Peroxidase
(EC 1.11.1.5)	Cytochrome c551 peroxidase
(EC 1.11.1.9)	Glutathione peroxidase

Superoxide Dismutases (SOD)

(EC 1.15.1.1)	Manganese superoxide dismutase
(EC 1.15.1.1)	Superoxide dismutase
(EC 1.15.1.1)	Superoxide dismutase [Cu-Zn] precursor
(EC 1.15.1.1)	Superoxide dismutase [Fe]
(EC 1.15.1.1)	Superoxide dismutase [Mn]
(EC 1.15.1.1)	Superoxide dismutase [Mn/Fe]
(EC 1.15.1.2)	Superoxide reductase

OTHERS (reduced rubredoxin + superoxide + 2 H+ = rubredoxin + H2O2)

Rubrerythrin

Rubredoxin

Thiamine biosynthesis (From purine pathway)

ThiC	phosphomethylpyrimidine synthase [EC:4.1.99.17]
THI20, ThiD	hydroxymethylpyrimidine/phosphomethylpyrimidine kinase / thiaminase [EC:2.7.1.49 2.7.4.7 3.5.99.2]
thiE	thiamine-phosphate pyrophosphorylase [EC:2.5.1.3]

Fermentations/Acetogens

Lactate to acetate	
(EC 1.1.2.3) (EC 1.1.1.27) (EC 1.1.1.28)	lactate dehydrogenase
(EC 2.3.1.54)	Pyruvate formate-lyase
(EC 6.2.1.13)	*Acetyl-CoA synthetase (ADP-forming)*
EC 2.3.1.8	*phosphate acetyltransferase*
EC 2.7.2.1	*acetate kinase*
Lactate to ethanol	
(EC 1.1.2.3) (EC 1.1.1.27) (EC 1.1.1.28)	lactate dehydrogenase
(EC 2.3.1.54)	Pyruvate formate-lyase
(EC 1.2.1.3)	Aldehyde dehydrogenase
(EC 1.1.1.1)	Alcohol dehydrogenase
Lactate to propionate	
(EC 6.4.1.1)	Pyruvate carboxyl transferase
(EC 1.1.1.37)	Malate dehydrogenase
(EC 4.2.1.2)	Fumarate hydratase
(EC 1.3.5.4)	succinate dehydrogenase
(EC 6.2.1.5)	Succinyl-CoA ligase [ADP-forming]
(EC 5.4.99.2)	Methylmalonyl-CoA mutase
(EC 4.1.1.41)	Methylmalonyl-CoA decarboxylase
(EC 6.2.1.1)	**Acetyl-coenzyme A synthetase**
Homoacetogens	
EC 6.3.4.3	formate:tetrahydrofolate ligase (ADP-forming)
EC 1.2.1.43	formate dehydrogenase
EC 3.5.4.9	5,10-methenyltetrahydrofolate 5-hydrolase
EC 1.5.1.20	methylenetetrahydrofolate reductase [NAD(P)H];

APPENDIX E

SUPPLEMENTARY MATERIAL FOR CHAPTER 6

Table E1: Sequencing statistics for SAR11 single cell amplified genomes.

DATASETS							NAR OPERON FRAGMENTS (#)				nitrate/nitrite transporter	COFACTOR BIOSYNTHESIS GENES	TERMINAL OXIDASES (gene Ids)		
SAG	Origin	N50	# contigs	total length (bp)	estimated total size (bp) based on genome completeness	Completeness	NARG	NARH	NARJ	NARI	NARK		BD-TYPE-subunit-I	BD-TYPE-subunit-II	COX1-type
A1_S1	GOM 1m	84,415	71	1,329,341	1,422,516	93	0	0	0	0					G56
A3_S3	GOM 1m	15,664	153	828,567	1,577,021	53	0	0	0	0					
A4_S4	GOM 1m	11,304	121	511,073	1,213,086	42	0	0	0	0					
A5_S5	GOM 1m	19,685	113	655,485	1,205,822	54	0	0	0	0					
A10_S10	ETNP 125m	5,349	127	347,621		38	1	2	0	0	1	moeA			G467,G468
A6_S6	ETNP 125m	8,646	147	551,754	1,243,529	44	1	1	1	1	2				
A7_S7	ETNP 125m	10,721	53	213,493		6	3	2	2	2					
A8_S8	ETNP 125m	3,153	83	170,771		7	0	0	0	0		moeA,mobA			
D3_S3	ETNP 125m	2,816	77	152,028		6	0	0	0	0					
B1_S11	ETNP 300m	2,744	205	385,868		21	0	0	0	0	2				
B2_S12	ETNP 300m	10,295	134	540,421		26	0	0	0	0	1	moeA,mobA			G407,G410
B3_S13	ETNP 300m	5,975	142	393,530	1,695,519	23	0	0	0	0					
B4_S14	ETNP 300m	2,876	93	186,991		6	0	0	0	0					
D10_S10	ETNP 300m	2,626	148	287,990		11	3	1	0	0			G211		G222
D9_S9	ETNP 300m	9,645	119	459,768	1,018,764	45	1	1	1	1			G342	G341	G376
E2_S13	ETNP 300m	6,134	108	376,244	1,150,242	33	1	1	1	0	2				
E3_S14	ETNP 300m	4,532	66	177,476		22	0	0	0	0	1	moeA			
E4_S15	ETNP 300m	9,809	100	381,935	1,537,581	25	1	2	0	0	1			G6	G218,G245
E5_S16	ETNP 300m	16,196	84	392,247	1,341,015	29	3	1	0	0	1	moeA,mobA			G420,G62

Table E2: Reference SAR11 genomes from cultured isolates.

Strain	Origin	Type	Clade	BioSample	Size (Mb)	Organism/Name
HTCC7217	Sargasso, BATS 10m	Isolate	Ia	SAMN02841150	1.43361	Candidatus Pelagibacter ubique HTCC7217
HTCC7214	Sargasso, BATS 10m	Isolate	Ia	SAMN02841172	1.37506	Candidatus Pelagibacter ubique HTCC7214
HTCC7211	Sargasso, BATS 10m	Isolate	Ia	SAMN02436224	1.45689	Candidatus Pelagibacter sp. HTCC7211
HTCC1062	Newport, Oregon	Isolate	Ia	SAMN02603690	1.30876	Candidatus Pelagibacter ubique HTCC1062
HTCC1040	Newport, Oregon	Isolate	Ia	SAMN02256395	1.27462	Candidatus Pelagibacter ubique HTCC1040
HTCC1016	Newport, Oregon	Isolate	Ia	SAMN02256429	1.2971	Candidatus Pelagibacter ubique HTCC1016
HTCC1013	Newport, Oregon	Isolate	Ia	SAMN02441456	1.30112	Candidatus Pelagibacter ubique HTCC1013
HTCC1002	Newport, Oregon	isolate	Ia	SAMN02436088	1.32862	Candidatus Pelagibacter ubique HTCC1002
HTCC8051	Newport, Oregon	isolate	Ia	SAMN02440710	1.39502	Candidatus Pelagibacter ubique HTCC8051
HIMB058	Kaneohe Bay	SAG	IIa.B *	SAMN02440920	1.11505	Candidatus Pelagibacter ubique HIMB058
HIMB5	Kaneohe Bay	isolate	Ia	SAMN00016662	1.3432	alpha proteobacterium HIMB5
HTCC9022		isolate	Ia	SAMN02440781	1.36096	Candidatus Pelagibacter ubique HTCC9022
HIMB083		isolate	Ia	SAMN02597166	1.396	Candidatus Pelagibacter ubique HIMB083
SCGC AAA240-E13	Aloha station 770m	SAG	Ic	SAMN02597172	1.4016	alpha proteobacterium SCGC AAA240-E13
SCGC AAA288-E13	Aloha station 770m	SAG	Ic	SAMN02597171	0.812863	alpha proteobacterium SCGC AAA288-E13
SCGC AAA288-G21	Aloha station 770m	SAG	Ic	SAMN02597281	0.909786	alpha proteobacterium SCGC AAA288-G21
SCGC AAA288-N07	Aloha station 770m	SAG	Ic	SAMN02597280	0.954664	alpha proteobacterium SCGC AAA288-N07
IMCC9063	Svalbard, Norway	Isolate	III	SAMN02603337	1.28473	Candidatus Pelagibacter sp. IMCC9063
HIMB114	Kaneohe Bay	isolate	III	SAMN02436217	1.23737	alpha proteobacterium HIMB114
HIMB59	Kaneohe Bay	isolate	V	SAMN00010387	1.41013	alpha proteobacterium HIMB59

Table E3: Oceanic metagenomic datasets and physicochemical parameters.

Key for references for each collection of metagenomic dataset:

A. Ganesh et al, ISME, 2015
B. Ganesh et al, ISME, 2014
C. This study
D. Coleman and Chisholm, PNAS, 2010
E. DeLong et al, Science, 2006
F. Konstantinidis et al. AEM, 2009
G. Eloe at al, PLOS One, 2011
H. Quaiser et al, ISME, 2011
I. Martín-Cuadrado et al, PLOS One, 2007
J. Ghai et al, ISME, 2010

	Sample/Dataset information			Sequencing Statistics			Biochemistry Data							
metagenome id	Ref.	Depth	SITE	platform	#Quality Trimmed reads. rRNA excluded for Transcriptomes	Mean read Length	DO µmol kg-1	NO3 µmol kg-1	NO2 µmol kg-1	Temp (˚C)	Salinity (PSU)	Sampling Sate	Latitude	Longitude
ETSP 70m		70m	ETSP	454	210,391	160	0.046			12.835	34.818	21/23/2010	20° 04.999 S	70° 48.001 W
ETSP 110m		110m	ETSP	454	194,730	137	0.022			12.265	34.836	21/23/2010	20° 04.999 S	70° 48.001 W
ETSP 200m		200m	ETSP	454	205,267	134	0.018			11.513	34.799	21/23/2010	20° 04.999 S	70° 48.001 W
ETSP 1000m	A	1000m	ETSP	454	115,988	152	57.990			4.535	34.508	11/23/10	20° 04.999 S	70° 48.001 W
ETNP 30m		30m	ETNP station 6	illumina	2,864,786	191	200.043	0	0	24.48	34.59	6/19/13	18° 54.0' N	104° 54.0' W
ETNP 85m		85m	ETNP	illumina	4,184,936	180	0.416	22.37	0.07	15.69	34.66	6/19/13	18° 54.0' N	104° 54.0' W
ETNP 100m		100m	ETNP	illumina	2,699,158	187	0.008	16.25	4.19	14.66	34.75	6/19/13	18° 54.0' N	104° 54.0' W
ETNP 125m		125m	ETNP	illumina	1,835,666	188	0.012	15.48	6.06	13.45	34.8	6/19/13	18° 54.0' N	104° 54.0' W
ETNP 300m	B	300m	ETNP	illumina	8,525,980	168	0.006	21.93	1.23	10.57	34.69	6/19/13	18° 54.0' N	104° 54.0' W
ETNP 68m (2014)		68m	ETNP	illumina	67,779,292	148	4.11-STOX**	38.79	0.07	18.51	34.59	5/19/14	18° 54' 3.1788"N	108° 48' 9.5394"
ETNP 120m (2014)	C	120m	ETNP	illumina	63,206,650	149	5.85 - CTD	12.91	0.04	13.54	34.71	5/19/14	18° 54' 3.1788"N	108° 48' 9.5394"
Transcriptome: ETNP 30m		30m	ETNP	illumina	272,020	156	200.043	0	0	24.48	34.59	6/19/13	18° 54.0' N	104° 54.0' W
Transcriptome: ETNP 85m		85m	ETNP	illumina	588,737	174	0.416	22.37	0.07	15.69	34.66	6/19/13	18° 54.0' N	104° 54.0' W
Transcriptome: ETNP 100m		100m	ETNP	illumina	188,677	179	0.008	16.25	4.19	14.66	34.75	6/19/13	18° 54.0' N	104° 54.0' W
Transcriptome: ETNP 125m		125m	ETNP	illumina	384,489	184	0.012	15.48	6.06	13.45	34.8	6/19/13	18° 54.0' N	104° 54.0' W
Transcriptome: ETNP 300m	A	300m	ETNP	illumina	174,627	179	0.006	21.93	1.23	10.57	34.69	6/19/13	18° 54.0' N	104° 54.0' W
GOM 1m		5m	GOM	illumina	27,734,948	141	193.340			27.18	36.224	5/29/12	26°00'21.1"N	92°21'57.8"W
GOM 25m		25m	GOM	illumina	30,001,998	142	200.535			26.15	36.360	5/29/12	26°00'21.1"N	92°21'57.8"W
GOM 73m		73m	GOM	illumina	27,279,280	141	175.898			21.17	36.501	5/29/12	26°00'21.1"N	92°21'57.8"W
GOM 150m		150m	GOM	illumina	23,753,160	139	114.706			15.86	36.078	5/29/12	26°00'21.1"N	92°21'57.8"W
GOM 300m		300m	GOM	illumina	19,354,890	141	105.864			11.46	35.410	5/29/12	26°00'21.1"N	92°21'57.8"W
GOM 600m		600	GOM	illumina	117,866,340	140	122.708			7.17	34.924	5/29/12	26°00'21.1"N	92°21'57.8"W
GOM 1000m		1000	GOM	illumina	26,429,176	138	171.699			4.99	34.926	5/29/12	26°00'21.1"N	92°21'57.8"W
GOM 1470m		1470	GOM	illumina	25,140,660	138	201.792			4.28	34.960	5/29/12	26°00'21.1"N	92°21'57.8"W
GOM 2107m	C	2300	GOM	illumina	27,934,226	139	206.194			4.25	34.965	5/29/12	26°00'21.1"N	92°21'57.8"W
BATS20.454		20m	BATS	454	327,870	229	204.900			26.7	36.44	10/1/06	31°40'01.2"N	64°10'01.2"W
BATS50.454		50m	BATS	454	402,923	248	205.400			24.0	36.74	10/1/06	31°40'01.2"N	64°10'01.2"W
BATS100.454	D	500m	BATS	454	476,890	222	210.800			19.6	36.69	10/1/06	31°40'01.2"N	64°10'01.2"W
HOT25.454		25m	HOT	454	351,469	108	205.000			26.4	35.08	03-2006	22°75'00.0"N	158°00'00.0"W
HOT75.454		125m	HOT	454	**320,698**	108	217.400			24.93	35.21	03-2006	22°75'00.0"N	158°00'00.0"W
HOT125.454		75m	HOT	454	366,501	110	204.900			22.19	35.31	03-2006	22°75'00.0"N	158°00'00.0"W
HOT500.454	E	500m	HOT	454	336,469	106	118.000	30.19		7.25	34.07	03-2006	22°75'00.0"N	158°00'00.0"W
HOT4000.sanger	F	4000m	HOT	fosmid ends	54,509	1421	147.800			1.46	34.69	12/21/03	22°75'00.0"N	158°00'00.0"W
PRT	G	6000m	PRT	454	445,714	405	240.000	24.2	0	NA	34.84		19°40'01.2"N	65°57'57.6"W
Marmara	H	1000m	Marmara	454	227,413	183	30.0	4.8		14.2	38.75	6/6/07	40°50'18.0"N	28°01'24.0"E
Med3000	I	3010m	Ionian Sea	fosmid ends	**9,048**	796	203.7			13.9	38.7	11/17/14	38°04'04.8"N	0°13'55.2"E
Med50	J	50m	Med.Sea	454	1,125,040	262	265.000			15.9	38.6	10/15/07	38° 4'6.64'' N	0° 13'55.18'' W